REMODELING OF THE CHROMATIN LANDSCAPE FOR TRANSCRIPTIONAL REGULATION IN ZEBRAFISH EARLY EMBRYO

by

Mengyao Tan

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Oncological Sciences

The University of Utah

December 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Mengyao Tan**

has been approved by the following supervisory committee members:

Bradley R. Cairns	, Chair	**10/11/2016** Date Approved
Barbara J. Graves	, Member	**10/11/2016** Date Approved
David J. Grunwald	, Member	**10/11/2016** Date Approved
Rodney Stewart	, Member	**10/11/2016** Date Approved
H. Joseph Yost	, Member	**10/11/2016** Date Approved

and by **Bradley R. Cairns**, Chair/Dean of

the Department/College/School of **Oncological Sciences**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Cells of the early vertebrate embryo are distinct in their ability to commit into any cell lineage. How the embryo acquires this remarkable plasticity from two terminally differentiated gametes remains largely unknown. The plasticity in early embryo relies on achieving a unique transcriptome, which is regulated at multiple levels - including chromatin accessibility at developmental enhancers and genes. To understand the global landscape of chromatin accessibility during early embryogenesis, we utilized zebrafish embryos and explored three aspects of chromatin regulation.

We first focused on the ATPase subunits (Brg1 and Brm) of SWI/SNF complexes, which are important regulators of chromatin accessibility and gene expression in all eukaryotes. To understand where they act in the genome, we profiled the occupancy of Brg1 and Brm by ChIP-seq at three early embryonic stages around major onset of zygotic genome activation. We observed the occupancy of Brg1 and Brm during early embryogenesis is highly dynamic. The promoters of key pluripotency factors and other developmental transcription factors are robustly occupied by Brg1 and Brm. Interestingly, Brg1, but not Brm, is highly correlated with active histone modifications. However, only Brm commonly occupies gene bodies, which is dependent on transcription elongation. This work suggests SWI/SNF complexes might play important roles during early embryogenesis, and also reveals distinct roles of Brg1 and Brm in early zebrafish development.

We then profiled the global landscape of accessible chromatin by ATAC-seq at three embryonic stages, as well as one differentiated tissue, adult liver. The data suggest chromatin accessibility increases during early embryogenesis. Here, $\sim$60% of open chromatin regions reside at genic regions and are highly enriched at promoters. Furthermore, many interesting candidate transcription factors are revealed based on motif analyses. Finally, ATAC-seq fragments with length of 120-220bp, together with MNase-seq date are used to profile nucleosome positioning. Our data determines nucleosome positioning during early embryogenesis, also discovered many interesting sequence characteristics involved in nucleosome positioning at various gene features.

In summary, this work has extensively investigated the dynamics of chromatin landscape and the role of chromatin remodelers during early zebrafish development, which allow the comprehensive understanding of the regulation during early embryogenesis.

To my greatest parents and dearest husband.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my mentor, Brad Carins. I could not have accomplished my Ph.D. training without his support and guidance. His vision, ideas and suggestions are essential to the success of my projects. He has also helped me to improve my skills on presentation and writing, and more importantly to become an independent scientist. I also want to thank my committee members, David Grunwald, Barbara Grave, Rodney Stewart and Joe Yost, for all the enlightening discussions and support. Special thanks to David Grunwald, for his faith in me and the encouragement over the years.

I also want to thank the lovely members in the Cairns Lab for making the lab a great place to work. Special thanks to Alisha Schlichter and Maggie Kasten, who are so kind and caring, and make the lab as a home to me. I want to thank other lab members for their help and stimulating discussions. Also to the former lab members, Magda Potok, Sue Hammoud, Jahnvi Pflueger, Christian Pflueger and Shan-Fu Wu for teaching me different techniques and help from many aspects. I would also like to thank my collaborator and friend, Candice Wike, for her help, support and the great feedback in writing. I would especially thank my close friends, Archana Yerra, Yixuan Guo and Zhizhou Ye, for their care and support and much wonderful and unforgettable memory outside of lab.

Other than the lab members, I am also grateful to have many friends and wonderful colleagues in the Hunstman Cancer Insitute and in the University of

Utah. Many thanks to Jessica Askin and Dee DalPonte for taking care of the graduate school requirements for me and making sure I am on track. I also want to thank Brian Dally, who has been bugged by me a lot but still helped me to get back my sequencing result as soon as possible. I also want to acknowledge the bioinformatics experts, especially Tim Parnell. I would not have acquired my bioinformatics skills without his detailed demonstrations and tremendous help. Lastly, many thanks to my table tennis and badminton buddies for so many wonderful and fun moments.

I want to thank my parents for their unconditional love, support and encouragement. It is very hard for us to be apart on two different continents. At those traditional festivals when other families get a reunion, my parents are by themselves, and we can only celebrate together over the internet. Confucius said: “While your parents are alive, do not journey afar. If a journey has to be made, your direction must be told.” My parents give me the freedom to take this long journey, and believe I will fulfill my potential. I love you, dad and mom.

Lastly, I want to thank my husband, Yuchen, for his support and love. It is very tough to be long-distance. However, his care, patience and understanding become the strongest reliance to me and keep me moving forward in the difficult times.

I had a great and unforgettable experience in graduate school. I am grateful for having many friends and wonderful moments. I also appreciate the frustrations and overwhelming times which make me stronger and confident that nothing can beat me.

CHAPTER 1

INTRODUCTION

During early embryogenesis, the vertebrate embryo reprograms the terminally-differentiated gametes and becomes totipotent, which allows the embryonic cells to adopt any cell fate. The ability to differentiate into all cell types except extraembryonic lineage is known as pluripotency. Elucidating how pluripotency is established and regulated is pivotal for our understanding of development, reprogramming, regeneration and tumorigenesis.

This fascinating plasticity relies on highly regulated transcription circuits. Specifically, genes required for pluripotency and early development are selectively expressed, while lineage-specification genes are poised in a repressed but competent state for later expression and development (Jaenisch & Young, 2008; MacArthur et al., 2009). The pluripotent transcriptome is coordinately modulated at multiple levels, including transcription factors binding, DNA methylation, histone modification and chromatin accessibility (Lessard & Crabtree, 2010; Burton & Torres-Padilla, 2014). Among these, chromatin accessibility is especially important due to its direct impact on the ability of transcription factors and co-regulators to bind within chromatin. This dissertation aims to provide insights into the regulation of pluripotency from a chromatin perspective, using the zebrafish embryo as a model. In principle, we hope to determine the genome-wide chromatin landscape

in early zebrafish development, and how it is regulated by chromatin remodelers, particularly SWI/SNF complexes. To answer these questions, it is important to understand the basics of chromatin; how does chromatin state affect transcription; how chromatin state, especially chromatin accessibility is regulated. I would also introduce chromatin remodelers, the central regulators of chromatin accessibility, with a focus on SWI/SNF complexes. Furthermore, what is known about chromatin state in early embryos, and how it is regulated will also be introduced.

1.1 Chromatin and chromatin accessibility

Genomic DNA is packaged into chromatin, which allows a tight control of gene expression and remarkable compaction of the genome. In most cell types, the basic unit of chromatin is nucleosome, which is composed of $\sim$147bp DNA and an octamer of four histones, H2A, H2B, H3, and H4 (Kornberg, 1974). Each nucleosome is separated by linker DNA, which is occupied by the fifth histone, H1. The linker DNA and H1 ensure the regularly spacing of nucleosome and enable the condensation of nucleosomes into higher-order chromatin structure (Kornberg, 1977; Luger et al., 1997).

Besides the canonical histones, histone variants can also be incorporated into nucleosomes. For example, H3.3 and CenH3 are variants to H3, and H2A.Z, H2A.X and macroH2A are variants to H2A. The histone compositions greatly impact the stability of nucleosomes. In addition, histone tails are subject to a variety of posttranslational modifications including methylation (mono-, di-, tri-methylation), acetylation, phosphorylation, ubiquitination, and recently discovered sumoylation and crotonylation (Tan et al., 2011). These histone modifications are differentially

distributed at euchromatin and heterochromatin, and are tightly related to gene expression (Li et al., 2007).

Another important concept about chromatin is chromatin accessibility. When DNA is wrapped around histones, it is inaccessible to many (but not all) DNA-binding regulatory proteins due to the tight interaction with histones, which is known as chromatin inaccessibility. However, during gene expression, DNA repair and replication, the DNA must be made accessible to various factors mediating these processes (Clapier & Cairns, 2009; Lessard & Crabtree, 2010) (Figure 1.1). Cooperatively, histone composition, histone modification, as well as chromatin accessibility contribute to a highly dynamic and complex chromatin state that is pivotal in essentially all genome-related processes.

1.2 Chromatin state and transcription regulation

Chromatin state has a profound impact on transcription regulation at multiple levels. First, chromatin accessibility, as it is defined, determines whether the promoters or enhancers are permissive for the binding of transcription factors and co-regulators, which leads to gene activation or repression based on the nature of the transcription factors.

Second, histone modifications, which act individually or cooperatively, are involved in gene activation, repression and poising (Li et al., 2007). Transcriptionally active genes often have H3K4me3 at the promoters, which can be directly bound by TAF3 in TFIID of the transcription machinery (Vermeulen et al., 2007; Lauberth et al., 2013). H3K4me3 is also capable of recruiting chromatin remodelers to dissemble nucleosome and to open the chromatin (Wysocka et al., 2006). Fur-

thermore, histone acetylation marks, such as H3K27ac, are also highly enriched at active promoters. Acetylation disrupts the interaction between DNA and histone and slightly increases the mobility of nucleosome (Ferreira et al., 2007). Moreover, acetylation can be recognized by bromodomains present in the subunits of chromatin remodelers (e.g. Brg1, Brm, PRDM1) (Filippakopoulos & Knapp, 2012), which then increase the chromatin accessibility. Of note, active promoters are also DNA hypomethylated, which allows the binding of transcript factors and other co-regulators (Tate & Bird, 1993). Besides promoters, transcribed genes also bear H3K36me3 at gene bodies, which are reported to recruit histone deacetylases, preventing RNA Pol II from initiating transcription within gene bodies (Carrozza et al., 2005; Joshi & Struhl, 2005). It is interesting to note that H3K36me3 has been found on promoters of silenced genes in zebrafish sperm, which suggests distinct roles of H3K36me3 in different cellular contexts (Wu et al., 2011).

In contrast, promoters of transcriptionally silent genes carry high DNA methylation, which prevents protein binding. Additionally, H3K27me3 and H3K9me3 are prevalent in repressed genes, which either antagonize active marks or condense the chromatin by recruiting specific chromatin remodelers (Li et al., 2007). Interestingly, in ES cells, a combination of H3K4me3 and H3K27me3 has been discovered specifically on development genes (Bernstein et al., 2006). The co-occurrence of active and repressive marks is termed ”bivalent marks” and has been identified later in multipotent stem cells (Cui et al., 2009), early embryos (Vastenhouw et al., 2010; Lindeman et al., 2011), and germline (Hammoud et al., 2009; Wu et al., 2011). Bivalent marks poise genes in a repressed but competent state for later expression and development.

The chromatin state at enhancers also plays a key role in regulating transcription. Enhancers are predominately marked by H3K4me1, which has been proposed to either prevents the *de novo* methylation at enhancers or promotes incorporation of H2A.Z to create a permissive chromatin. This in turn facilitates TF binding and leads to the expression of target genes (Ooi et al., 2007; Altaf et al., 2010). H3K4me1 in combination of H3K27ac is enriched at active enhancers (Calo & Wysocka, 2013), whereas co-localization of H3K4me1 and H3K27me3 is an indicator of poised enhancers (Rada-Iglesias et al., 2011). To sum up, transcription is tightly regulated at the chromatin level, which suggests chromatin regulation may play a key role in the establishment and maintenance of pluripotency in early embryo, as pluripotency relies on highly regulated transcription circuits.

1.3 Regulators of chromatin state and SWI/SNF complexes

The diverse chromatin states are regulated by a variety of proteins and complexes, which can be generally divided into two classes, histone modifiers and chromatin remodelers. Histone modifiers include ”histone writers” and ”histone erasers” (Eberharter & Becker, 2002). For instance, histone acetyltransferases (p300 or CBP) establish H3K27ac, which can be removed by histone deacetylases (HDACs). Chromatin remodelers are multisubunit protein complexes that use the energy from ATP hydrolysis to alter nucleosome compositions or slide/eject nucleosomes. They can be divided into four major subfamilies – SWI/SNF, ISWI, INO80 and CHD – based on the domain structures of the ATPase subunits, subunits, complex subunit composition, and their specialized functions (Clapier & Cairns, 2009). In general, chromatin remodelers that assemble chromatin by organizing

nucleosome arrays are involved in gene repression, whereas chromatin remodelers associated with chromatin/nucleosome disorganization and disassembly promote gene activation. Among these, SWI/SNF complexes are well-studied for their involvement in gene activation via disorganizing chromatin, which is a focus of my dissertation.

The complex was initially discovered in yeast (Winston & Carlson, 1992; Cairns et al., 1996), and similar complexes were later identified in fly and vertebrates (Mohrmann & Verrijzer, 2005). In yeast, there are two classes of SWI/SNF family chromatin remodelers, one is RSC complexes, the other is SWI/SNF complexes. The RSC complexes is about ten-fold more abundant than SWI/SNF complexes and and is essential for yeast viability (Cairns et al., 1996). In vertebrates, two major SWI/SNF complexes exist: BAF (Brg1-assoicated factor) and PBAF (polybromo Brg1-assoicated factor). The two complexes are comprised of one of two mutually exclusive ATPase subunits (Brg1 and Brm), several highly conserved core subunits (SNF5, BAF155 and BAF170), and some subunits that are unique to BAF or PBAF (Figure 1.2). Specifically, ARID1A and ARID1B are two mutually exclusive subunits of BAF complex, whereas PBRM1 and ARID2 are specific to PBAF complex (Wilson & Roberts, 2011). Other than the ubiquitously expressed BAF and PBAF complexes, tissue-specific SWI/SNF complexes have been widely identified, which consist of tissue-specific subunits. For instance, ES cell-specific BAF complex (esBAF) is characterized by the replacement of BAF170 subunits with a second copy of BAF155 subunit (Ho et al., 2009b). Additionally, nBAF, another well-studied BAF complex, is specifically present in postmitotic neurons, and is distinct in its incorporation of BAF53b, BAF45b and CREST which are

exclusively expressed in neurons (Wu et al., 2007).

1.4 ATPase subunits of SWI/SNF complex

Brg1 and Brm are two mutually exclusive ATPase subunits of SWI/SNF complexes, conserved in vertebrates. The two proteins are very similar in protein sequence, where their protein sequences share 72% identities in zebrafish and 78% identities in human. Why do vertebrates need two ATPase subunits if they are so similar in protein sequence? Is it because Brg1 and Brm have specialized functions? Expression profiles reveal that Brg1 and Brm are co-expressed in various cells and tissues (Reyes et al., 1998; Becker et al., 2009). However, Brg1 is preferentially expressed in cells that undergo proliferation and self-renewal, whereas Brm appears more abundant in cells that are not constantly proliferating (Reisman et al., 2005). Moreover, Brg1 interacts with zinc finger proteins (KLF, GATA, etc.) through a unique N'terminal domain which is absent in Brm. On the other hand, only Brm binds ankyrin repeat proteins (ICD22 and CBF-1), which are the components of Notch signaling pathways (Kadam & Emerson, 2003). In contrast to the binding specificity, Brg1 and Brm has also been shown to interact with retinoblastoma (Rb) protein, glucocorticoid receptor and C/EBP (Muchardt & Yaniv, 1993; Kowenz-Leutz & Leutz, 1999).

In order to better elucidate the specialized functions of Brg1 and Brm, a few studies directly compared the roles of Brg1 and Brm in different systems. In osteoblast differentiation, Brm deletion accelerates differentiation, whereas Brg1 deletion inhibits differentiation, suggesting antagonistic roles of Brg1 and Brm (Flowers et al., 2009). In skeletal myogenesis, both Brg1 and Brm are required for

MyoD-mediated differentiation (de la Serna et al., 2001). However, Brg1 functions on gene activation at early stages of differentiation, whereas Brm promotes Ccnd1 repression and cell cycle arrest prior to the activation of muscle genes (Albini et al., 2015). In addition, my data suggest Brg1 and Brm might have shared functions, as well as specialized roles. Collectively, these data suggest that Brg1 and Brm may have similar, auxiliary and antagonistic roles depending on the cellular contexts. Nevertheless, the most evident differences between Brg1 and Brm are from genetics studies in mouse and cancer studies, which will be discussed in the following two sections.

1.5 SWI/SNF complex and early embryo development

To become a totipotent zygote, the embryos need to reprogram two differentiated gametes via extensive chromatin reorganization. Chromatin remodelers, especially SWI/SNF complexes, regulate chromatin accessibility and are important in the early development. In mice, zygotic knockdown of Brg1 renders embryonic lethality at preimplantation stages, while maternal-zygotic Brg1 null mouse embryo is arrested at as early as the two-cell stage (which is when ZGA happen in mouse) (Bultman et al., 2000, 2006). In line with its role in gene activation, 30% of genes are downregulated in Brg1 maternal null mouse, and exhibit a decrease of H3K4me2. In contrast, zygotic Brm null mouse is viable and fertile, and only exhibits mild defects including increased body weight and deregulation of cell growth control in embryonic fibroblasts (Reisman et al., 2005). However, it has to be acknowledged that the Brm null mouse can circumvent the Brm deletion by alternative splicing to create a short isoform containing all essential domains

(Thompson et al., 2015). Therefore, complete knockout of Brm is needed to compare the importance of Brg1 and Brm directly. Besides the ATPase subunits, inactivations of other core subunits (ARID1A, ARID1B, BAF155 and SNF5) also result in embryonic lethality. Furthermore, deletion of Brm, which is the only ATPase subunit of SWI/SNF complex in fly, results in severe embryonic defects and eventfully leads to lethality before adulthood (Brizuela et al., 1994).

Additional support came from studies in ES cells. It has been demonstrated that the esBAF complex is required for maintaining the pluripotency of mouse ES cells (Ho et al., 2009b). Moreover, SWI/SNF complexes bind pluripotent genes, and displayed extensive genome-wide colocalization with Oct4, Nanog and Sox2 (Ho et al., 2009a; Kidder et al., 2009). Knockdown of Brg1 in mouse ES cell downregulated pluripotency related genes, while upregulated the genes involved in differentiation (Kidder et al., 2009). Furthermore, in human ES cell, Brg1 has been found localized at a large number of enhancers (Rada-Iglesias et al., 2011). These evidence, suggest SWI/SNF complexes, represented by Brg1, may increase promoter and enhancer accessibility, and govern the proper transcription regulation in ES cells.

1.6 SWI/SNF complexes and tumorigenesis

With the extensive high-throughput sequencing of cancer genomes, SWI/SNF complexes have emerged as major tumor suppressors, as more than 20% of human tumors contain mutations on subunits of SWI/SNF complexes (Kadoch et al., 2013)(Figure 1.3). Around 400 mutations have been found in Brg1 among a variety of tumors. Most of the mutations are high-impact missense substitutions or

truncating mutations. In contrast, only around 100 mutations are identified in Brm, which is still a high mutation rate (Hodges et al., 2016). SNF5 is another subunit that is mutated in nearly all rhabdoid tumors. Conditional biallelic inactivation of SNF5 in mouse renders cancer phenotype as early as 11 weeks, which is even more aggressive compared to p53 deficiency where cancer onsets around 20 weeks (Roberts et al., 2002). In addition, ARID1A and PRDM1 are frequently mutated in a number of cancers. ARID1A mutations were reported in $\sim$50% of ovarian clear cell carcinomas, $\sim$34% of uterine endometrial carcinomas, $\sim$29% of bladder tumors and $\sim$34% stomach tumors (Hodges et al., 2016). PRDM1 mutations or losses are identified in 41% of renal clear cell carcinomas (Varela et al., 2011). All these reported mutations make SWI/SNF complexes the most frequently mutated chromatin regulatory complexes in cancer. However, the mechanism yet remains obscure. Therefore, understanding the functions of SWI/SNF complexes in both normal development and cancer contexts will provide valuable insights into our understanding on tumorigenesis and therapeutic approaches.

1.7 Chromatin dynamics in mouse early development

Zygotic genome activation (ZGA) is one of the most important molecular events in early embryogenesis, when the zygotic genome transits from transcriptional silence to transcriptional activation (Figure 1.4). At ZGA, genes that are required for early development are selectively expressed, while genes needed for lineage specification and organogenesis are primed in a repressed state. This plasticity is achieved in part through coordinate regulation of chromatin accessibility.

Histone modifications are highly dynamic in the early embryos and display

parental asymmetries. Early studies in mouse embryos using immunofluorescence and mass spectrometry revealed unique histone modifications profiles. Upon fertilization, the maternal genome displays a histone modification profile largely resembling somatic cells, where discernable levels of acetylated lysine, methylated H3K4me, H3K9me2/3, H3K20me3 and H3K64me3 are detected. H3K20me3 and H3K64me3 are then rapidly removed at 2-cell stage (Adenot et al., 1997; Kourmouli et al., 2004; Santos et al., 2005). For paternal genome, exchange of protamines to hyperacetylated histones supplied from the oocytes initiates immediately after fertilization. Then the acetylated histones are substituted with primarily monomethylated histones (i.e. H3K4me, H3K9me, H3K27me, and H3K20me) (Kimmins & Sassone-Corsi, 2005; Lepikhov & Walter, 2004). In addition, H3.3 is incorporated into the paternal genome, which is important for pericentric heterochromatin formation (Santenard et al., 2010).

Very recently, three groups overcame technical challenges with early mouse embryo, and performed ChIP-seq on H3K4me3, H3K27ac, and H3K27me in gametes and pre-implantation embryos (Liu et al., 2016; Zhang et al., 2016; Dahl et al., 2016). All three studies found that H3K4me3 displays a broad distribution specifically in oocyte, which becomes confined to narrow distribution at two-cell stage, as seen in ES cells and differentiated cells. Although they detected co-occurrence of H3K4me3 and H3K27me3 at certain genes in early two-cell embryos, it is infrequent and unstable (Liu et al., 2016).

Finally, the genome-wide chromatin accessibility in early embryos has begun to be uncovered by liDNase or ATAC-seq experiments (Lu et al., 2016; Wu et al., 2016). These data revealed that the chromatin is less accessible before the genome

activation and becomes increasingly open after ZGA, which is in contrast to the result from electron-microscopy (EM) where the zygote chromatin is less condensed compared to that of later-stage embryo (Ahmed et al., 2010). The authors argued that the EM results reflect a less-established higher-order chromatin structure but not a permissive local chromatin state. Regardless, the genome-wide profiling of open chromatin revealed the accessible regions usually located at promoters and especially at enhancers. Additionally, the accessible regions changes extensively at different stage of embryogenesis. In summary, dynamic chromatin reprogramming and establishment take place in early development, which is essential for pluripotency and cell fate specification.

1.8 The advantage of zebrafish as a model organism

Zebrafish has been extensively used as an important model organism in the studies on vertebrate development, regeneration, and diseases (Dooley & Zon, 2000). An adult zebrafish is around 3 to 4 centimeters long, and can be maintained in large scale easily and economically. One pair of male and female zebrafish can produce around 200 embryos per week. The embryos fertilize and develop outside of the mother, which enables easy collection of thousands of embryos in one day. In addition, the zebrafish embryos are nearly transparent, which allows a clear visualization of reporter genes and the examination of phenotype. Moreover, the zebrafish embryo develops rapidly: it establishes most cell types and tissues by five days postfertilization, and reaches sexual maturity around 2 to 3 months (Kimmel et al., 1995). These features make zebrafish suitable for large-scale forward genetic screen and promote the development of various means for re-

verse genetics such as MO knockdown, ZFNs, TALENS and CPISPR technology (Lawson & Wolfe, 2011; Zu et al., 2013; Auer et al., 2014).

Furthermore, the zebrafish genome is well annotated and composed of around 26,000 protein-coding genes, most of which are highly conserved (Howe et al., 2013). Notably, many key chromatin regulators and transcription factors regulating early embryogenesis are conserved and function in comparable pathways (Mudbhary & Sadler, 2011). Importantly, zygotic genome activation in zebrafish happens at 1000-cell stage around 3hpf (hour postfertilization), while ZGA happens at 2-cell stage in mouse, and at 2-cell to 4-cell stage in humans (Figure 1.4). The abundance of cells at ZGA and the feasible collection of a vast number of embryos make zebrafish a great system for studying early embryogenesis.

1.9 Regulation of early zebrafish embryogenesis

In zebrafish, transcription factors, small RNAs and chromatin regulators are central players in early embryogenesis. For the proper transition from maternal to zygotic control of embryogenesis, maternal factors have to be destabilized in a regulatory way. In zebrafish, miR-430 directly modulates the degradation of several hundred maternal provided mRNA targets, which is in part via deadenylation (Giraldez et al., 2006). Deficiency of miR-430 in *dicer* null fish (Dicer is the enzyme required for microRNA biogenesis) leads to defects in gastrulation and delayed development, suggesting a vital role of miRNA in the early embryogenesis (Giraldez et al., 2005). Here, a key question is which factors help to promote the early and robust transcription of miR-430, which is informed in this dissertation.

Besides transcription factors and small RNAs, DNA methylation is also im-

portant in the early development. Inhibiting DNMTs (DNA methyltransferases) by 5-Aza-2'-deoxycytidine arrested the embryos during gastrulation, and misregulated transcription of about 3,000 genes (Martin et al., 1999; Potok et al., 2013). Consistently, knockdown of DNMT1 causes intestinal and pancreatic differentiation abnormalities, and knockdown of DNMT3 leads to neurogenic defects (Rai et al., 2006; Anderson et al., 2009; Rai et al., 2010).

Dynamic histone modification profiles are also observed in zebrafish embryos. Before ZGA, low levels of H3K4me3 is detected on around a thousand of the genes that are required for early development, such as pou5f1. After ZGA, the levels of histone modifications increase dramatically, where more than 80% of genes acquire H3K4me3 at the promoters (Vastenhouw et al., 2010). Bivalent marks are also established and poise the development genes (i.e. Hox genes) for later expression (Lindeman et al., 2011; Wu et al., 2011). Furthermore, some zebrafish HDAC (histone deacetylases) mutants display abnormalities in early development and various tissue developments (Mudbhary & Sadler, 2011). Although mutants of most histone modifiers have not been established in zebrafish, it is still likely that histone modifiers play indispensable roles in zebrafish early development, considering the conservation of histone modifiers between zebrafish and mammals, and the important roles of histone modifiers in mouse embryogenesis (Lessard & Crabtree, 2010).

Pou5f1, Nanog, and Sox2 are core pluripotency factors of stem cell self-renewal. Their conserved zebrafish homologs are also involved in the establishment of pluripotency in early embryo. Based on genetic inactivation of Pou5f1 (annotated as Pou5f3 in zebrafish genome build, zv10) and morpholino inhibition of Nanog

and SoxB1 factor (Sox2, Sox2, Sox19a, and Sox19b), the Lee group found that the treated embryo is completely arrested before gastrulation, accompanied by a substantial gene downregulation at ZGA (Lee et al., 2013). Meanwhile, another group demonstrated Pou5f1 preferentially binds around 40% of genes that are activated at the ZGA, and probably enhances their gene expression. Of note, the transcription at ZGA is downregulated but not abolished when simultaneously inactivated Pou5f1, Nanog, and SoxB1 factors, suggesting other unknown factors are also required for gene activation at ZGA (Leichsenring et al., 2013). One intriguing question is how did Pou5f1, Nanog, and SoxB1 factors manage to bind the promoter of their target genes. Are SWI/SNF complexes involved in augmenting the local chromatin accessibility, thus facilitating their binding? Our data suggest so, and will be described in detail later in Chapter 2.

In conclusion, the studies on small RNAs, chromatin regulators and transcription factors are valuable for our understanding of the regulation of embryogenesis. However, two major aspects of embryogenesis regulation are still poorly characterized in zebrafish: 1) the role of chromatin remodelers, and 2) the genome-wide chromatin accessibility. Understanding the chromatin landscape and how it is shaped by chromatin remodelers will provide invaluable insights to our knowledge of early embryogenesis, cell plasticity and reprogramming, which is the focus of my dissertation.

1.10 Dissertation overview

In this dissertation, works from three major projects are highlighted. These projects include the first genome-wide profiling of Brg1 and Brm in the early ze-

brafish embryo and the dynamic chromatin landscape in early embryos and adult liver based on ATAC-seq and MNase-seq.

In Chapter 2, ChIP-seq data of Brg1 and Brm at embryonic stages are described. We revealed the occupancy of Brg1 and Brm in early embryos is highly dynamic. Brg1 and Brm co-occupy a large number of genes, but also bind specific genes, respectively. In addition, Brg1 and Brm occupancy correlates with higher gene expression, more accessible chromatin and active histone modifications. One distinct feature of Brg1 is the correlation with active histone modifications (H3K27ac, H3K4me1 and H3K4me3) but not with repressive histone modification, H3K27me3. Uniquely for Brm, it occupies not only promoters but also gene bodies of a large number of genes, which perfectly correlates with the occupancy of Pol II. We further demonstrated the distribution of Brm at gene body is dependent on transcription elongation, suggesting coordination of functions.

In Chapter 3, extensive bioinformatics analyses of ATAC-seq from three embryonic stages and adult liver are presented. We revealed the chromatin accessibility at Pol II genes and enhancers increases substantially from pre-MTB to MBT. A large set of open chromatin regions are present at genic regions, with a preference at promoters. Genes that are associated with open chromatin are generally expressed, except a group of genes that are marked by H3K4me1 and H3K27me3 at their promoters. My analyses also extend to Pol III genes, and show the number of Pol III genes which are associated with open chromatin decreases as embryo develops $-$ a trend opposite to Pol II. In addition, motif analysis is performed to identify transcription factors that are potentially involved in the regulation of early embryogenesis. I confirmed the enrichment of known factors including Pou5f3

and Sox-family factors, and also identified novel motifs, such as Nr3e2 and NFYB. Finally, fragments with length of 120-220bp, which presumably represent mono-nucleosomes, can be utilized to profile nucleosome positioning.

In Chapter 4, I focus on nucleosome positioning at embryonic stages and adult liver, which is profiled by MNase-seq. Nucleosome positioning from crosslinked and noncrosslinked samples are largely similar. However, crosslinked samples have a better representation of 'weak' nucleosomes compared to noncrosslinked samples. Comparisons of nucleosome positioning at different stages suggest embryonic samples at pre-MBT are distinct from that of later stages, especially at gene or promoter regions. Additionally, the establishment of nucleosome array is not dependent on transcription, but may be related to the poised Pol II and DNA sequences. Furthermore, nucleosomes are also enriched on exons, which could be mediated by the high frequency of polyA(5) sequence at the 5'end of exons and the alternation of A/T and G/C dinucleotides inside exons. Interestingly, exons with strong AT-GC alternation often have weak polyA(5) sequence at their 5end, whereas exons with weak AT-GC alternation tend to have strong polyA(5) sequence – a feature that is also observed in other organisms.

In Appendix A, I summarize all the commercial antibodies that have been tested for various zebrafish transcription factors and chromatin remodelers. A couple of the antibodies robustly detect their target proteins, and a few have been validated for their application in ChIP-seq. However, a lot of the tested antibodies fail to recognize their target proteins, which are also listed to help others avoid wasting time on them.

In Appendix B, I summarize an additional 10 customized antibodies that have

been generated but not used in previous projects. The detailed information regarding antigen regions, cloning primers, and peptide purification are described. Four antibodies displayed promising results based on Western blotting, including Oct1, SNF2H, SNF2L and Sox3.

Lastly, in Appendix C, plasmids containing cDNA of two transcription factors and four chromatin remodelers have been constructed. Dominant negative constructs of Brg1, Brm, SNF2H have also been generated.

1.11 References

Adenot, P.G., Mercier, Y., Renard, J.P., Thompson, E.M. (1997). Differential h4 acetylation of paternal and maternal chromatin precedes dna replication and differential transcriptional activity in pronuclei of 1-cell mouse embryos. Development *124*, 4615–4625.

Ahmed, K., Dehghani, H., Rugg-Gunn, P., Fussner, E., Rossant, J., Bazett-Jones, D.P. (2010). Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. PloS One *5*, e10531.

Albini, S., Toto, P.C., Dall'Agnese, A., Malecova, B., Cenciarelli, C., Felsani, A., Caruso, M., Bultman, S.J., Puri, P.L. (2015). Brahma is required for cell cycle arrest and late muscle gene expression during skeletal myogenesis. EMBO Reports , e201540159.

Altaf, M., Auger, A., Monnet-Saksouk, J., Brodeur, J., Piquet, S., Cramet, M., Bouchard, N., Lacoste, N., Utley, R.T., Gaudreau, L., et al. (2010). Nua4-dependent acetylation of nucleosomal histones h4 and h2a directly stimulates incorporation of h2a. z by the swr1 complex. Journal of Biological Chemistry *285*, 15966–15977.

Anderson, R.M., Bosch, J.A., Goll, M.G., Hesselson, D., Dong, P.D.S., Shin, D., Chi, N.C., Shin, C.H., Schlegel, A., Halpern, M., et al. (2009). Loss of dnmt1 catalytic activity reveals multiple roles for dna methylation during pancreas development and regeneration. Developmental Biology *334*, 213–223.

Auer, T.O., Duroure, K., De Cian, A., Concordet, J.P., Del Bene, F. (2014). Highly efficient crispr/cas9-mediated knock-in in zebrafish by homology-independent dna repair. Genome Research *24*, 142–153.

Becker, T.M., Haferkamp, S., Dijkstra, M.K., Scurr, L.L., Frausto, M., Diefenbach, E., Scolyer, R.A., Reisman, D.N., Mann, G.J., Kefford, R.F., et al. (2009). The chromatin remodelling factor brg1 is a novel binding partner of the tumor suppressor p16 ink4a. Molecular Cancer *8*, 1.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Brizuela, B.J., Elfring, L., Ballard, J., Tamkun, J.W., Kennison, J.A. (1994). Genetic analysis of the brahma gene of drosophila melanogaster and polytene chromosome subdivisions 72ab. Genetics *137*, 803–813.

Bultman, S.J., Gebuhr, T.C., Pan, H., Svoboda, P., Schultz, R.M., Magnuson, T. (2006). Maternal brg1 regulates zygotic genome activation in the mouse. Genes & Development *20*, 1744–1754.

Bultman, S., Gebuhr, T., Yee, D., La Mantia, C., Nicholson, J., Gilliam, A., Randazzo, F., Metzger, D., Chambon, P., Crabtree, G., et al. (2000). A brg1 null mutation in the mouse reveals functional differences among mammalian swi/snf complexes. Molecular Cell *6*, 1287–1295.

Burton, A., Torres-Padilla, M.E. (2014). Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. Nature Reviews Molecular Cell Biology *15*, 723–735.

Cairns, B.R., Lorch, Y., Li, Y., Zhang, M., Lacomis, L., Erdjument-Bromage, H., Tempst, P., Du, J., Laurent, B., Kornberg, R.D. (1996). Rsc, an essential, abundant chromatin-remodeling complex. Cell *87*, 1249–1260.

Calo, E., Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Molecular Cell *49*, 825–837.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P., et al. (2005). Histone h3 methylation by set2 directs deacetylation of coding regions by rpd3s to suppress spurious intragenic transcription. Cell *123*, 581–592.

Clapier, C.R., Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annual Review of Biochemistry *78*, 273–304.

Cui, K., Zang, C., Roh, T.Y., Schones, D.E., Childs, R.W., Peng, W., Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell *4*, 80–93.

Dahl, J.A., Jung, I., Aanes, H., Greggains, G.D., Manaf, A., Lerdrup, M., Li, G., Kuan, S., Li, B., Lee, A.Y., et al. (2016). Broad histone h3k4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. Nature .

Dooley, K., Zon, L.I. (2000). Zebrafish: a model system for the study of human disease. Current Opinion in Genetics & Development *10*, 252–256.

Eberharter, A., Becker, P.B. (2002). Histone acetylation: a switch between repressive and permissive chromatin. EMBO Reports *3*, 224–229.

Ferreira, H., Flaus, A., Owen-Hughes, T. (2007). Histone modifications influence the action of snf2 family remodelling enzymes by different mechanisms. Journal of Molecular Biology *374*, 563–579.

Filippakopoulos, P., Knapp, S. (2012). The bromodomain interaction module. FEBS Letters *586*, 2692–2704.

Flowers, S., Nagl, N.G., Beck, G.R., Moran, E. (2009). Antagonistic roles for brm and brg1 swi/snf complexes in differentiation. Journal of Biological Chemistry *284*, 10067–10075.

Giraldez, A.J., Cinalli, R.M., Glasner, M.E., Enright, A.J., Thomson, J.M., Baskerville, S., Hammond, S.M., Bartel, D.P., Schier, A.F. (2005). Microrna regulate brain morphogenesis in zebrafish. Science *308*, 833–838.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., Schier, A.F. (2006). Zebrafish mir-430 promotes deadenylation and clearance of maternal mrnas. Science *312*, 75–79.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478.

Ho, L., Jothi, R., Ronan, J.L., Cui, K., Zhao, K., Crabtree, G.R. (2009a). An embryonic stem cell chromatin remodeling complex, esbaf, is an essential component of the core pluripotency transcriptional network. Proceedings of the National Academy of Sciences *106*, 5187–5191.

Ho, L., Ronan, J.L., Wu, J., Staahl, B.T., Chen, L., Kuo, A., Lessard, J., Nesvizhskii, A.I., Ranish, J., Crabtree, G.R. (2009b). An embryonic stem cell chromatin remodeling complex, esbaf, is essential for embryonic stem cell self-renewal and pluripotency. Proceedings of the National Academy of Sciences *106*, 5181–5186.

Hodges, C., Kirkland, J.G., Crabtree, G.R. (2016). The many roles of baf (mswi/snf) and pbaf complexes in cancer. Cold Spring Harbor Perspectives in Medicine *6*, a026930.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498–503.

Jaenisch, R., Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567–582.

Joshi, A.A., Struhl, K. (2005). Eaf3 chromodomain interaction with methylated h3-k36 links histone deacetylation to pol ii elongation. Molecular Cell *20*, 971–978.

Kadam, S., Emerson, B.M. (2003). Transcriptional specificity of human swi/snf brg1 and brm chromatin remodeling complexes. Molecular Cell *11*, 377–389.

Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J., Crabtree, G.R. (2013). Proteomic and bioinformatic analysis of mammalian swi/snf complexes identifies extensive roles in human malignancy. Nature Genetics *45*, 592–601.

Kidder, B.L., Palmer, S., Knott, J.G. (2009). Swi/snf-brg1 regulates self-renewal and occupies core pluripotency-related genes in embryonic stem cells. Stem Cells *27*, 317–328.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Developmental Dynamics *203*, 253–310.

Kimmins, S., Sassone-Corsi, P. (2005). Chromatin remodelling and epigenetic features of germ cells. Nature *434*, 583–589.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and dna. Science *184*, 868–871.

Kornberg, R.D. (1977). Structure of chromatin. Annual Review of Biochemistry *46*, 931–954.

Kourmouli, N., Jeppesen, P., Mahadevhaiah, S., Burgoyne, P., Wu, R., Gilbert, D.M., Bongiorni, S., Prantera, G., Fanti, L., Pimpinelli, S., et al. (2004). Heterochromatin and tri-methylated lysine 20 of histone h4 in animals. Journal of Cell Science *117*, 2491–2501.

Kowenz-Leutz, E., Leutz, A. (1999). Ac/ebpβ isoform recruits the swi/snf complex to activate myeloid genes. Molecular Cell *4*, 735–743.

Lauberth, S.M., Nakayama, T., Wu, X., Ferris, A.L., Tang, Z., Hughes, S.H., Roeder, R.G. (2013). H3k4me3 interactions with taf3 regulate preinitiation complex assembly and selective gene activation. Cell *152*, 1021–1036.

Lawson, N.D., Wolfe, S.A. (2011). Forward and reverse genetic approaches for the analysis of vertebrate development in the zebrafish. Developmental Cell *21*, 48–64.

Lee, M.T., Bonneau, A.R., Takacs, C.M., Bazzini, A.A., DiVito, K.R., Fleming, E.S., Giraldez, A.J. (2013). Nanog, pou5f1 and soxb1 activate zygotic gene expression during the maternal-to-zygotic transition. Nature *503*, 360–364.

Leichsenring, M., Maes, J., Mössner, R., Driever, W., Onichtchouk, D. (2013). Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science *341*, 1005–1009.

Lepikhov, K., Walter, J. (2004). Differential dynamics of histone h3 methylation at positions k4 and k9 in the mouse zygote. BMC Developmental Biology *4*, 1.

Lessard, J.A., Crabtree, G.R. (2010). Chromatin regulatory mechanisms in pluripotency. Annual Review of Cell and Developmental Biology *26*, 503.

Li, B., Carey, M., Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Lindeman, L.C., Andersen, I.S., Reiner, A.H., Li, N., Aanes, H., Østrup, O., Winata, C., Mathavan, S., Müller, F., Aleström, P., et al. (2011). Prepatterning of developmental gene expression by modified histones before zygotic genome activation. Developmental Cell *21*, 993–1004.

Liu, X., Wang, C., Liu, W., Li, J., Li, C., Kou, X., Chen, J., Zhao, Y., Gao, H., Wang, H., et al. (2016). Distinct features of h3k4me3 and h3k27me3 chromatin domains in pre-implantation embryos. Nature .

Lu, F., Liu, Y., Inoue, A., Suzuki, T., Zhao, K., Zhang, Y. (2016). Establishing chromatin regulatory landscape during mouse preimplantation development. Cell *165*, 1375–1388.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 å resolution. Nature *389*, 251–260.

MacArthur, B.D., Ma'ayan, A., Lemischka, I.R. (2009). Systems biology of stem cell fate and cellular reprogramming. Nature Reviews Molecular Cell Biology *10*, 672–681.

Martin, C.C., Laforest, L., Akimenko, M.A., Ekker, M. (1999). A role for dna methylation in gastrulation and somite patterning. Developmental Biology *206*, 189–205.

Mohrmann, L., Verrijzer, C.P. (2005). Composition and functional specificity of swi2/snf2 class chromatin remodeling complexes. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression *1681*, 59–73.

Muchardt, C., Yaniv, M. (1993). A human homologue of saccharomyces cerevisiae snf2/swi2 and drosophila brm genes potentiates transcriptional activation by the glucocorticoid receptor. The EMBO Journal *12*, 4279.

Mudbhary, R., Sadler, K.C. (2011). Epigenetics, development, and cancer: zebrafish make their mark. Birth Defects Research Part C: Embryo Today: Reviews *93*, 194–203.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., et al. (2007). Dnmt3l connects unmethylated lysine 4 of histone h3 to de novo methylation of dna. Nature *448*, 714–717.

Potok, M.E., Nix, D.A., Parnell, T.J., Cairns, B.R. (2013). Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell *153*, 759–772.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rai, K., Jafri, I.F., Chidester, S., James, S.R., Karpf, A.R., Cairns, B.R., Jones, D.A. (2010). Dnmt3 and g9a cooperate for tissue-specific development in zebrafish. Journal of Biological Chemistry *285*, 4110–4121.

Rai, K., Nadauld, L.D., Chidester, S., Manos, E.J., James, S.R., Karpf, A.R., Cairns, B.R., Jones, D.A. (2006). Zebra fish dnmt1 and suv39h1 regulate organ-specific terminal differentiation during development. Molecular and Cellular Biology *26*, 7077–7085.

Reisman, D.N., Sciarrotta, J., Bouldin, T.W., Weissman, B.E., Funkhouser, W.K. (2005). The expression of the swi/snf atpase subunits brg1 and brm in normal human tissues. Applied Immunohistochemistry & Molecular Morphology *13*, 66–74.

Reyes, J., Barra, J., Muchardt, C., Camus, A., Babinet, C., Yaniv, M. (1998). Altered control of cellular proliferation in the absence of mammalian brahma (snf2α). The EMBO Journal *17*, 6979–6991.

Roberts, C.W., Leroux, M.M., Fleming, M.D., Orkin, S.H. (2002). Highly penetrant, rapid tumorigenesis through conditional inversion of the tumor suppressor gene snf5. Cancer Cell *2*, 415–425.

Santenard, A., Ziegler-Birling, C., Koch, M., Tora, L., Bannister, A.J., Torres-Padilla, M.E. (2010). Heterochromatin formation in the mouse embryo requires critical residues of the histone variant h3. 3. Nature Cell Biology *12*, 853–862.

Santos, F., Peters, A.H., Otte, A.P., Reik, W., Dean, W. (2005). Dynamic chromatin modifications characterise the first cell cycle in mouse embryos. Developmental Biology *280*, 225–236.

de la Serna, I.L., Carlson, K.A., Imbalzano, A.N. (2001). Mammalian swi/snf complexes promote myod-mediated muscle differentiation. Nature Genetics *27*.

Tadros, W., Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. Development *136*, 3033–3042.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell *146*, 1016–1028.

Tate, P.H., Bird, A.P. (1993). Effects of dna methylation on dna-binding proteins and gene expression. Current Opinion in Genetics & Development *3*, 226–231.

Thompson, K.W., Marquez, S.B., Lu, L., Reisman, D. (2015). Induction of functional brm protein from brm knockout mice. Oncoscience *2*, 349.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C.K., Stephens, P., Davies, H., Jones, D., Lin, M.L., Teague, J., et al. (2011). Exome sequencing identifies frequent mutation of the swi/snf complex gene pbrm1 in renal carcinoma. Nature *469*, 539–542.

Vastenhouw, N.L., Zhang, Y., Woods, I.G., Imam, F., Regev, A., Liu, X.S., Rinn, J., Schier, A.F. (2010). Chromatin signature of embryonic pluripotency is established during genome activation. Nature *464*, 922–926.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.P., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., Timmers, H.T.M. (2007). Selective anchoring of tfiid to nucleosomes by trimethylation of histone h3 lysine 4. Cell *131*, 58–69.

Wilson, B.G., Roberts, C.W. (2011). Swi/snf nucleosome remodellers and cancer. Nature Reviews Cancer *11*, 481–492.

Winston, F., Carlson, M. (1992). Yeast snf/swi transcriptional activators and the spt/sin chromatin connection. Trends in Genetics *8*, 387–391.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., et al. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature *534*, 652–657.

Wu, J.I., Lessard, J., Olave, I.A., Qiu, Z., Ghosh, A., Graef, I.A., Crabtree, G.R. (2007). Regulation of dendritic development by neuron-specific chromatin remodeling complexes. Neuron *56*, 94–108.

Wu, S.F., Zhang, H., Cairns, B.R. (2011). Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm. Genome Research *21*, 578–589.

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., et al. (2006). A phd finger of nurf couples histone h3 lysine 4 trimethylation with chromatin remodelling. Nature *442*, 86–90.

Zhang, B., Zheng, H., Huang, B., Li, W., Xiang, Y., Peng, X., Ming, J., Wu, X., Zhang, Y., Xu, Q., et al. (2016). Allelic reprogramming of the histone modification h3k4me3 in early mammalian development. Nature *537*, 553–557.

Zhou, V.W., Goren, A., Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nature Reviews Genetics *12*, 7–18.

Zu, Y., Tong, X., Wang, Z., Liu, D., Pan, R., Li, Z., Hu, Y., Luo, Z., Huang, P., Wu, Q., et al. (2013). Talen-mediated precise genome modification by homologous recombination in zebrafish. Nature Methods *10*, 329–331.

Figure 1.1: Epigenetic regulation of transcription. Simplified inactive gene is on the left and active gene is on the right. From the top to bottom, status of DNA methylation, histone modifications and histone variants, chromatin accessibility, and higher-order structures such as nuclear lamina-associated domains are depicted. Genes that are inactive have methylated DNA and inaccessible chromatin. They are also associated with repressive histone marks, H3K9me3 and H3K27me3, as well as lamina. In contrast, active genes have DNA hypomethylation (or absence), accessible chromatin. They are also associated with active histone marks, H3K4me3 and H3K36me3. In addition, they reside in transcription factory domain. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Zhou et al., 2011)

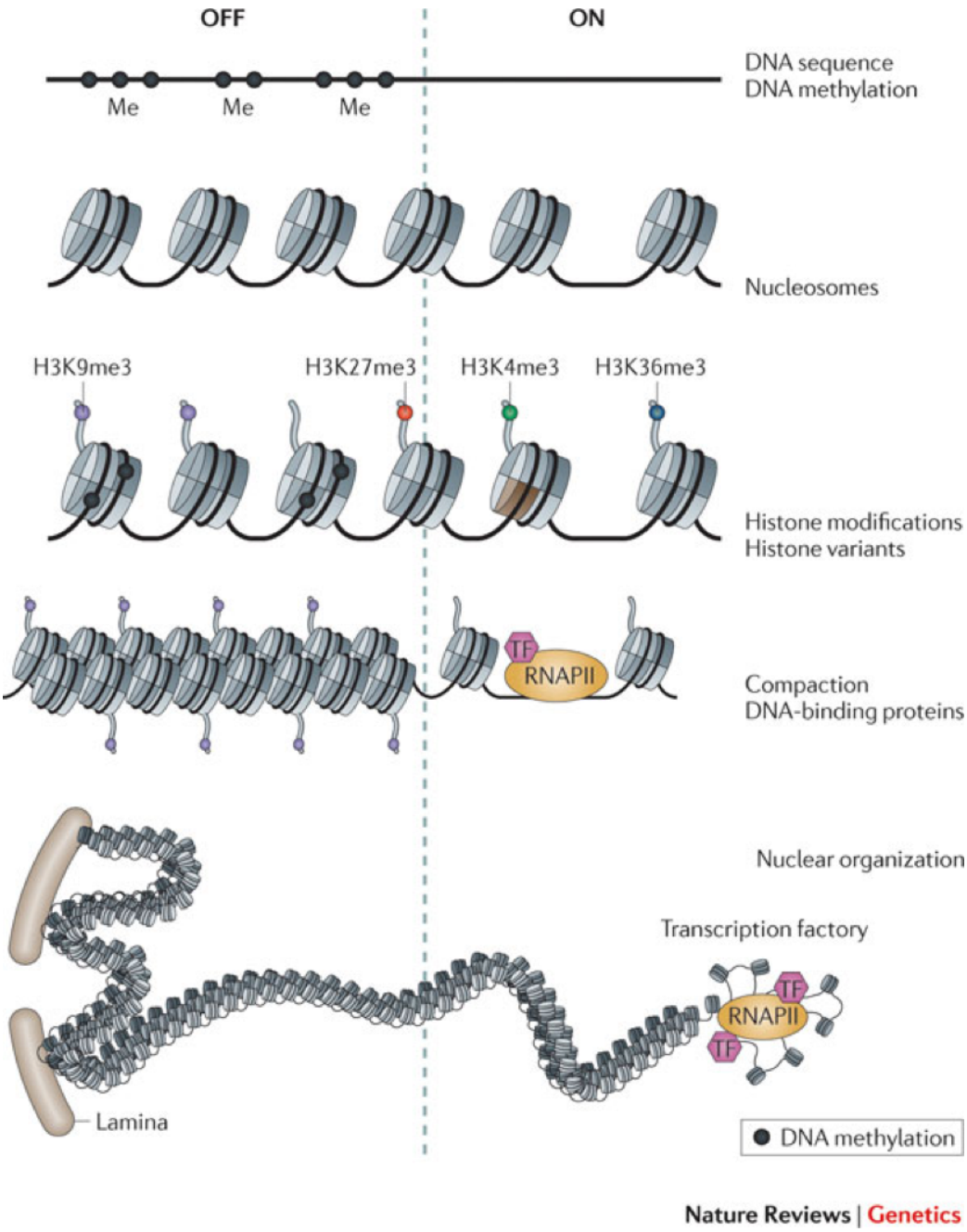
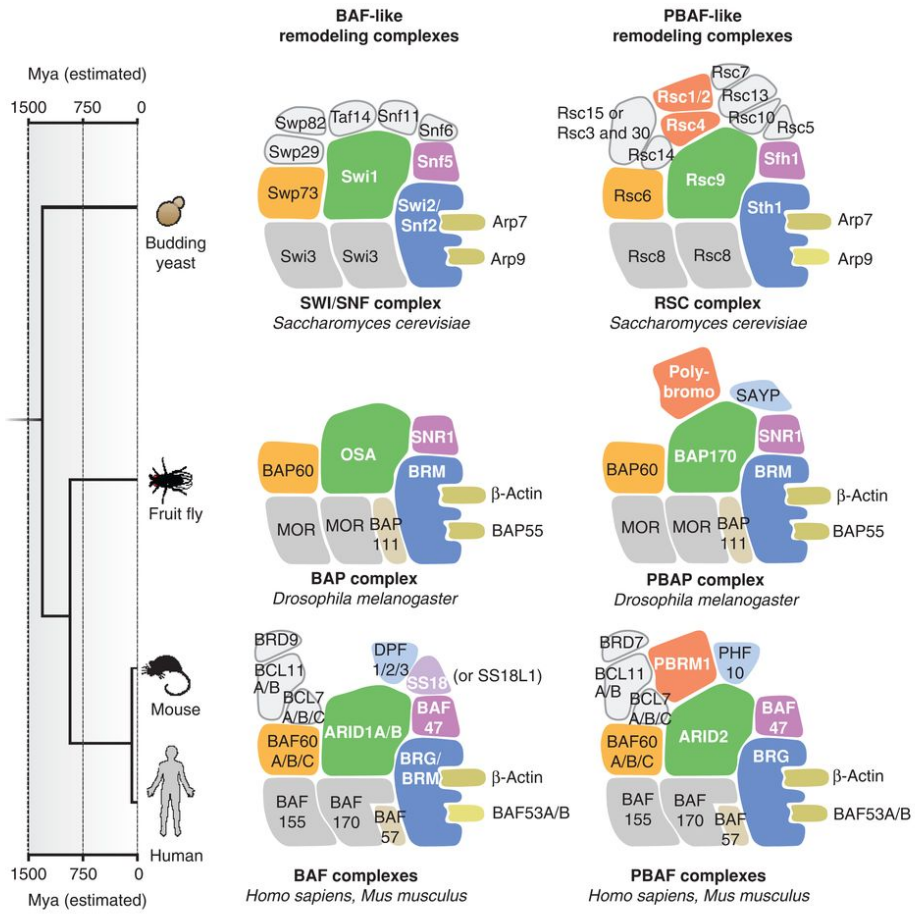

OFF
ON
DNA sequence
DNA methylation
Me
Me
Me
Nucleosomes
H3K9me3
H3K27me3
H3K4me3
H3K36me3
Histone modifications
Histone variants
TF
RNAPII
Compaction
DNA-binding proteins
Nuclear organization
Transcription factory
TF
RNAPII
TF
Lamina
DNA methylation
Nature Reviews | Genetics

Figure 1.2: Homology of SWI/SNF complexes in different organisms. (A) BAF- and PBAF- like complexes in yeast (SWI/SNF and RSC complexes), fly (Brahma-associated proteins (BAPs) and Polybromo-associated BAP (PBAP) complexes), mouse and human (BAF and PBAF complexes). (B) Summary of the alternative names of subunits of SWI/SNF complexes. Reprinted by permission of Cold Spring Harbor Laboratory Press (Hodges et al., 2016)

A



B

HUGO symbol	Alternative names	HUGO symbol	Alternative names
ACTL6A/B	BAF53A/B	**SMARCA2**	BRM
ARID1A/B	BAF250A/B	**SMARCA4**	BRG
ARID2	BAF200	**SMARCB1**	BAF47, SNF5, INI1
BCL7A/B/C	–	**SMARCC1/2**	BAF155/BAF170
BCL11A/B	–	**SMARCD1/2/3**	BAF60A/B/C
BRD7/9	–	**SMARCE1**	BAF57
DPF1/2/3	BAF45B/D/C	**SS18**	–
PBRM1	BAF180, Polybromo	**SS18L1**	CREST
PHF10	BAF45A		

Figure 1.3: Mutation frequency of SWI/SNF complexes subunits in cancer. Frequency includes all nonsilent mutations, biallelic deletions, and gene fusions. MRT, malignant rhabdoid tumor; PCNSL, Primary central nervous system lymphoma; DLBCL, diffuse large B-cell lymphoma; AML, acute myeloid leukemia. Reprinted by permission of Cold Spring Harbor Laboratory Press, adapted from (Hodges et al., 2016)

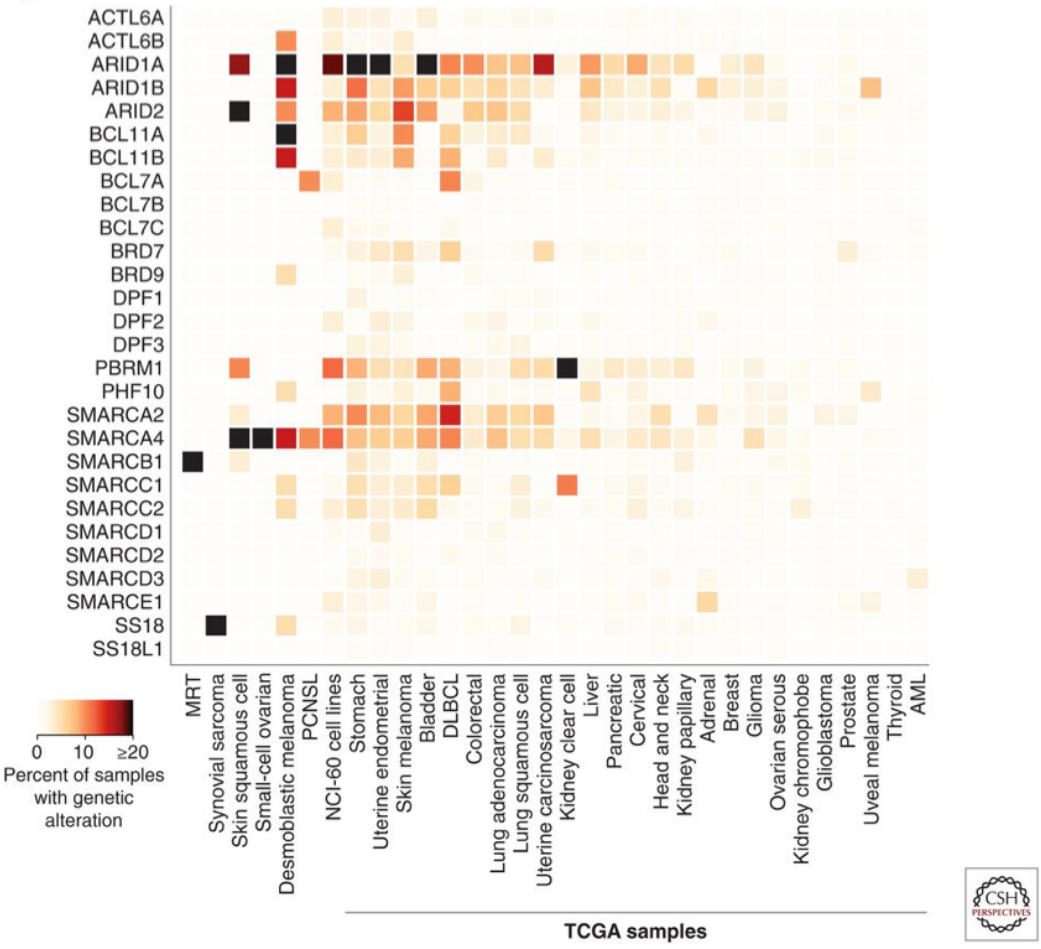
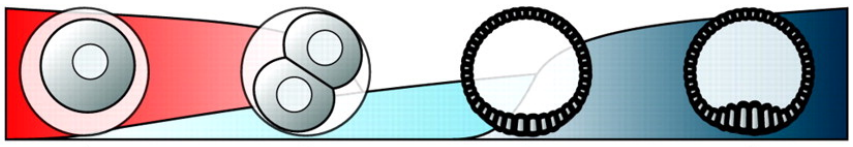

ACTL6A
ACTL6B
ARID1A
ARID1B
ARID2
BCL11A
BCL11B
BCL7A
BCL7B
BCL7C
BRD7
BRD9
DPF1
DPF2
DPF3
PBRM1
PHF10
SMARCA2
SMARCA4
SMARCB1
SMARCC1
SMARCC2
SMARCD1
SMARCD2
SMARCD3
SMARCE1
SS18
SS18L1
0
10
≥20
Percent of samples with genetic alteration
MRT
Synovial sarcoma
Skin squamous cell
Small-cell ovarian
Desmoblastic melanoma
PCNSL
NCI-60 cell lines
Stomach
Uterine endometrial
Skin melanoma
Bladder
DLBCL
Colorectal
Lung adenocarcinoma
Lung squamous cell
Uterine carcinosarcoma
Kidney clear cell
Liver
Pancreatic
Cervical
Head and neck
Kidney papillary
Adrenal
Breast
Glioma
Ovarian serous
Kidney chromophobe
Glioblastoma
Prostate
Uveal melanoma
Thyroid
AML
TCGA samples

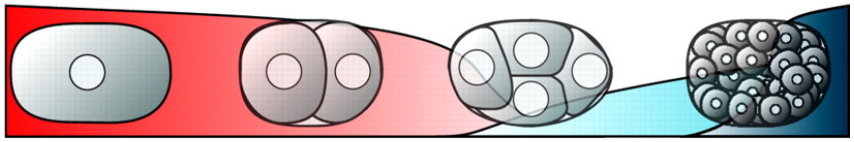
Figure 1.4: Overview of maternal-zygotic transition in different organisms. Red zone indicates the degradation of maternal provided factors; light blue zone indicates the 'minor' wave of zygotic genome activation (ZGA); dark blue zone indicates the 'major' wave of ZGA. Cell cycles and respective time for key embryonic stages are labelled under each organisms. Reprinted by permission of THE COMPANY OF BIOLOGISTS LTD. (Tadros & Lipshitz, 2009)

Sea urchin (*S. purpuratus*)



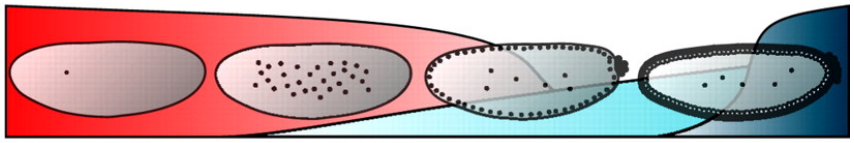
Cleavage cycle	1	2	~8	~10
Time (hours)	0	2	15	30

Nematode (*C. elegans*)



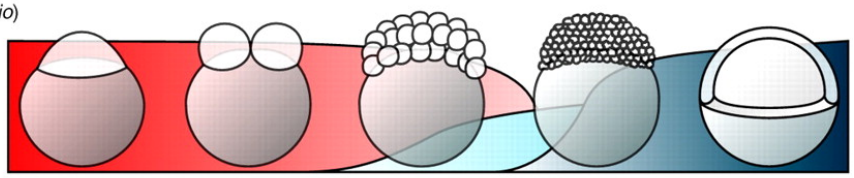
Cleavage cycle	0	1	2	6/7
Time (hours)	0	0.7	1.25	3-3.5

Fruit fly (*D. melanogaster*)



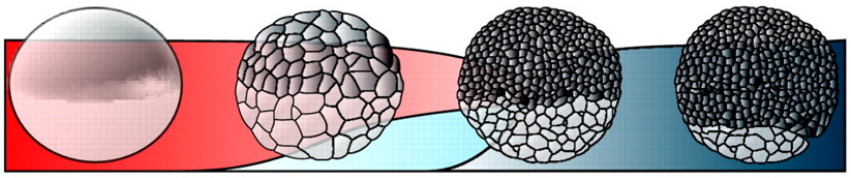
Cleavage cycle	0	8	10	14
Time (hours)	0	1	1.5	2.5

Zebrafish (*D. rerio*)



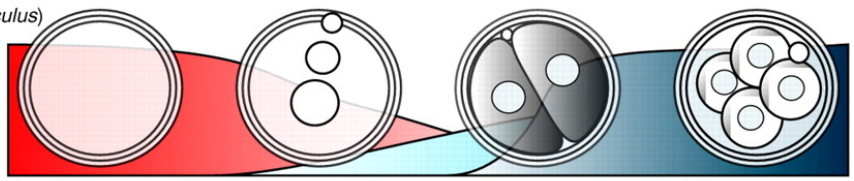
Cleavage cycle	0	1	6	10	14
Time (hours)	0	0.75	2	2.75	5.25

Frog (*X. laevis*)



Cleavage cycle	0	6	13	14
Time (hours)	0	4	5	9

Mouse (*M. musculus*)



Cleavage cycle	0	0	1	2
Time (hours)	0	10	22	37

CHAPTER 2

DISTINCTIVE ROLES OF BRG1 AND BRM IN EARLY ZEBRAFISH DEVELOPMENT

2.1 Abstract

Brg1 and Brm function as alternative core enzymatic ATPases of SWI/SNF complex in vertebrates, with clear homologs in zebrafish. Brg1 (but not Brm) is required for early cleavage-stage transcription in mice, but where and how Brg1 or Brm impact chromatin in early embryos remains unexplored. Here, we utilize zebrafish embryos to provide the first simultaneous genome-wide profiling of Brg1 and Brm occupancy in any organism/cell type, alongside our profiling of open chromatin and RNA polymerase II. We reveal shared and unique loci for Brg1 and Brm, their shared occupancy of open chromatin and particular developmental promoters, and higher transcription at co-occupied genes. Interestingly, only Brg1 is strongly associated with active histone modifications. Strikingly, only Brm commonly occupies gene bodies, overlapping precisely with Pol II. Functional experiments involving Pol II elongation inhibition alongside bioinformatic analyses suggest that Brm travels with Pol II into gene bodies, primarily at short, very highly-transcribed genes with few introns. Taken together, our results support roles for Brg1 primarily at promoters and enhancers, and for Brm in association with Pol II in elongation through chromatin within gene bodies during genome-wide

transcriptional activation.

2.2 Introduction

Early embryogenesis is a fascinating stage in development, where the embryo – formed from two highly specialized terminally-differential gametes – gains the ability to differentiate into any cell lineage. These pluripotent embryonic cells have to selectively express genes required for pluripotency and early development, while also poising lineage-specification genes in a repressed but competent state for later expression and development (Jaenisch & Young, 2008; Zernicka-Goetz et al., 2009; Lee et al., 2014). This tight control is established through precise regulation of chromatin state at multiple levels, including chromatin accessibility – which involves interplay between transcription factors and chromatin remodeling complexes (Clapier & Cairns, 2009).

Here, we explore the roles of SWI/SNF subfamily chromatin remodelers in very early development, using the advantages of the zebrafish model. SWI/SNF complexes are known to play important roles in helping transcription factors bind to chromatin. They are multi-subunit chromatin remodeling complexes that are conserved in eukaryotes, and consist of a highly-conserved set of six core subunits accompanied by an additional set of auxiliary subunits, which can vary between organisms and/or cell types (Clapier & Cairns, 2009; Hota & Bruneau, 2016). Among the core subunits is a central catalytic ATPase, which functions as an ATP-dependent DNA translocase, which can pump DNA around nucleosomes. SWI/SNF complexes promote accessibility of transcription factors to chromatin through utilizing their DNA translocation activity to slide or eject nucleosomes,

and can facilitate the binding of either activators or repressors; therefore, they can facilitate either activation or repression – though activation appears more common (Clapier & Cairns, 2009).

In vertebrates, SWI/SNF-subfamily complexes are built around one of two core ATPase subunits: Brg1 or Brm, which are highly similar, and assemble in a mutually-exclusive manner into the complex. Brg1 and Brm can interact with distinct classes of transcription factors, and thus potentially localize to different genomic loci – however, as no genome-wide study of Brm occupancy currently exists, its localization remains largely unknown (Kadam & Emerson, 2003). *Brg1* knockout confers embryonic lethality at preimplantation stages, whereas the *Brm* mutant mouse appears viable and fertile, though 15% heavier than controls (Bultman et al., 2000; Reyes et al., 1998). However, alternative splicing in certain cell types may allow the current *Brm* insertion mouse to splice around the insertion to create a slightly smaller functional isoform containing all essential domains – requiring a full deletion/knockout to definitively determine the *Brm* null phenotype and contribution to cancer (Thompson et al., 2015). In addition, human *BRG1* and *BRM* are mutated in a number of malignancies, with *BRG1* more frequently disrupted than *BRM* (Hodges et al., 2016). Curiously, Brm but not Brg1 appears involved in alternative splicing, with Brm occupancy demonstrated at three test genes (Batsché et al., 2006).

A key question in very early development is how the genome transitions from transcriptional quiescence (and reliance on maternally-derived proteins and mRNA) to transcriptional activation, termed zygotic genome activation (ZGA). Here, a major role for chromatin remodeling can be envisioned, to help sculpt the nucleosome

landscape and open chromatin for transcription factors to define genes that will be activated or repressed when transcription is enabled. However, assessing very early roles in the embryo for factors is challenging due to the maternal inheritance of protein and/or mRNA (requiring oocyte-specific mutant/depletion for genetic examination), and the limitations in cell numbers at ZGA, limiting genomics/molecular approaches. In mice and humans, major ZGA occurs at/around the 2-cell or 4-cell stage, respectively. However, in zebrafish major ZGA occurs around the tenth cell cycle ($\sim$1000 cells), enabling genomics approaches (Tadros & Lipshitz, 2009). In zebrafish, the larger process commonly referred to as the midblastula transition (termed MBT, cell cycles 10-12) is coincident with strongest phase of ZGA, so we will hereafter use the term MBT. Notably, elegant prior work in mice demonstrated that oocyte-specific loss of *Brg1* conferred a 2-cell arrest, strongly suggesting Brg1 involvement in ZGA (Bultman et al., 2006). However where Brg1 (or Brm) remodels chromatin and its relationship to RNA polymerases has not been determined. Beyond ZGA, disruption of particular core subunits of SWI/SNF complexes (e.g. Brg1, SNF5, or BAF155) in ES cells impairs pluripotency (Lessard & Crabtree, 2010). Mechanistically, SWI/SNF complexes occupy the promoters of core pluripotency factors (Oct4, Nanog, Sox2) and a subset of their target genes, which promotes the activation of pluripotency circuitry in ES cells (Ho et al., 2009). However, whether and how SWI/SNF complexes bind or regulate pluripotency genes in early embryos – mouse, human or zebrafish – remains unknown, as do the unique and redundant roles of Brg1 and Brm. In zebrafish, Brg1 and Brm proteins are highly similar to their orthologs in mice and humans (Table 2.1), and zebrafish genome contains orthologs of all core SWI/SNF subunits found in the

mouse and human genomes, as well as many of the accessory subunits (Table 2.1). Although oocyte-specific deletions of zebrafish *brg1* or *brm* have not been created, loss of *brg1* causes defects in early eye and heart development, and lethality at 6 days postfertilization (Eroglu et al., 2006), whereas no study has examined the roles of *brm* in zebrafish.

To understand how Brg1 and Brm regulate chromatin and transcription during early embryogenesis, including whether Brg1 and Brm have specialized functions at this stage, we mapped Brg1 and Brm genome occupancy in early zebrafish embryos, defined open chromatin regions via ATAC-seq, identified RNA polymerase II (Pol II)-occupied genes – and also performed functional experiments involving Pol II transcription inhibitors. Here, we reveal the dynamic properties and specialization of Brg1 and Brm during MBT stages, showing that Brg1 functions primarily at active promoters and enhancers, whereas Brm uniquely functions with Pol II, residing within the bodies of mainly short highly-transcribed genes during genome-wide transcriptional activation.

2.3 Results

2.3.1 Specific antibodies for Brg1 and Brm reveal embryonic expression

Rabbit polyclonal antibodies that specifically recognize zebrafish Brg1 or Brm were produced by raising antibodies against unique regions of zebrafish Brg1 and Brm proteins, and conducting affinity purification (Figure 2.S1A and 1B). Each purified antibody specifically recognized Brg1 or Brm, and displayed low or negligible cross-reactivity, respectively (Figure 2.S2A). Immunoblotting identified Brg1 at $\sim$200kDa and Brm at $\sim$170kDa, slightly larger (or smaller, respectively) than

their predicted sizes of 182kDa or 178kDa (Figure 2.S1C). We detected low levels of Brg1 and Brm proteins in oocytes and 1-cell embryos – suggesting maternal inheritance – and higher levels as early development progressed, with Brg1 protein levels moderately higher than Brm (Figure 2.S2B, C), consistent with mice (LeGouy et al., 1998).

2.3.2 Genomic profiling reveals Brg1- and Brm-specific regions during early development

To help understand the roles of Brg1 and Brm during early development, we performed ChIP-seq of Brg1 and Brm at pre-MBT (2.5hpf), MBT (4hpf) and post-MBT (5.3hpf) stages utilizing the two polyclonal antibodies (Figure 2.1A, Table 2.1). As the replicates (typically two) were highly similar (pairwise $r>0.95$) the datasets were combined and peaks were called by MACS2 (q-value FDR 0.01). For Brg1, we identified 94 occupied regions at pre-MBT, which increased substantially to 5,469 regions at MBT, and to 11,778 regions at post-MBT (Figure 2.1B). A similar trend was observed for Brm: 551 at pre-MBT, 11,482 regions at MBT, and 15,456 regions at post-MBT (Figure 2.1B).

To characterize these occupied regions, we examined their genomic distribution on different genomic features, as annotated by Refseq. First, we found that Brg1 and Brm are highly enriched at promoters (Figure 2.1C). Interestingly, Brm displayed a much higher fractional occupancy in coding (exon/intron) and downstream regions than did Brg1 (Figure 2.1C), attributes explored in more detail later. Examination of Brg1- and Brm-occupied genes at each stage reveals a moderate-high degree of overlap, along with a substantial number of genes occupied specifically by either Brg1 or Brm – a proportion which increases as development progresses

(Figure 2.1D). Notably, Gene Ontology (GO) analysis of Brg1 and Brm occupied genes at pre-MBT revealed 'transcription' and 'development' as the most enriched terms. At MBT, enrichment additionally included metabolic processes such as 'ribosome' and 'nucleotide binding', whereas post-MBT enriched terms like 'cell migration' and 'embryonic organ development' were prevalent (Figure 2.S3A).

To reveal how occupied regions change during development, we examined regions that gain or lose occupancy by Brg1 and Brm. To display, we provide Brg1 and Brm ChIP-seq signal at each stage aligned to the center of the occupied regions (Figure 2.2A, B). Occupied regions were separated into three categories: 'pre-MBT' (regions occupied pre-MBT), 'MBT gained' and 'post-MBT gained' – separately, for both Brg1 and Brm. Interestingly, the vast majority ($\sim$80%) of Brg1-occupied regions at pre-MBT were maintained at MBT (Figure 2.2A, B, Figure 2.S4A), and 81.5% of these Brg1-occupied regions at MBT were maintained post-MBT (Figure 2.2A, B). We observed a similar pattern for Brm, though a slightly lower portion of sites maintained occupancy during each developmental phase (Figure 2.2C, D, Figure 2.S4B). Consistent with our analysis of occupied regions, genes occupied by Brg1 and Brm at early stages were largely maintained at later stages (Figure 2.S4C).

To display these different modes of occupancy dynamics, we provide a gallery of genome snapshots (Figure 2D), which includes regions/genes consistently occupied (e.g. *cxcl12a* and *etv4*), those restricted to pre-MBT (Figure 2.2D *klf1, hoxb8b, zgc:66455*), 'MBT gained' (Figure 2.2D *bmp2b*) or 'post-MBT gained' (Figure 2.2D *tcf7*). In addition, we also identified regions specifically bound by Brg1 or Brm, which also have diverse occupancy dynamics (Figure 2.S5). Curiously,

sites that exhibit a 'switch' in occupancy, either Brm-to-Brg1 or vice versa, were very rare. Taken together, Brg1 and Brm genes increase in number from moderate ($<$200) pre-MBT to numerous ($>$5200) by post-MBT, and although they largely overlap, there are clear specific sites for Brm at all stages, and for Brg1 at all stages after pre-MBT.

2.3.3 Motif analyses and ChIP-seq profiling reveal candidate factors for SWI/SNF recruitment

To identify candidate transcription factors for SWI/SNF recruitment (Brg1- or Brm-containing) we employed Regulatory Sequence Analysis Tools (RSAT) (Medina-Rivera et al., 2015) to determine the over-represented motifs at Brg1- and Brm-occupied regions at three stages (Figure 2.3A). At pre-MBT, we found Rfx1, Eomes and Pou3f2 motifs enriched at both Brg1- and Brm-occupied regions, and Dmrt1 and HoxD motifs only at Brm-occupied. Later in development, motifs for Sox3, HoxD, and Runx1 were enriched, accompanied by the loss of Rfx1 and Dmrt3 motifs at both Brg1- and Brm-occupied regions. In addition, three unknown motifs (CAAAACA, CTGCAG and TACTGTA) were obtained. To determine whether the candidate recruiting motifs/co-factors for SWI/SNF complex at enhancers and promoter are different, we partitioned occupied regions into promoters and intergenic areas, and performed the same motif analysis (Figure 2.S6). This additionally revealed Sox and Dmrt motifs enriched at pre-MBT Brg1-occupied promoters, and motifs for Tbx5, Pou3f2 and unknown motif GCTAAC in intergenic regions.

Among the enriched motifs, Eomes, Pou3f2, and Sox3 are of high interest in very early development; Eomes promotes trophectoderm differentiation in mammals, while Pou5f1 and Sox factors promote pluripotency and ES-cell self-renewal,

and early zebrafish development (Russ et al., 2000; Niwa, 2007; Leichsenring et al., 2013). Beyond motif analysis, to assess co-occupancy of these factors with Brg1 or Brm, we processed the public ChIP-seq data of Eomesa (ortholog of mammalian Eomes) (Nelson et al., 2014), Sox2, Pou5f3 (annotated as Pou5f3 in the zebrafish genome build zv10 and as Pou5f1 in zebrafish genome build zv9, ortholog of mammalian Pou5f1) (Leichsenring et al., 2013) and Nanog (Xu et al., 2012). Interestingly, all four factors were highly enriched at the centers of Brg1-occupied regions at MBT, and moderately enriched at Brm-occupied regions (Figure 2.3B). Likewise, regions with high levels of Brg1 or Brm had high levels of these transcription factors, displayed either as heat maps (Figure 2.3C, Figure 2.S7) or quantiles based on q-value (Figure 2.S8) – and is observed at both promoters and distal regulatory elements (Figure 2.4D, Figure 2.S9). Interestingly, SWI/SNF complex and several TFs (e.g. Pou5f3, Sox2 and Nanog) bind themselves; each is co-localized at the promoter and distal regulatory elements of Pou5f3, Sox3, Nanog – whereas the Eomesa promoter/enhancer is not occupied. In sum, these data unveil a potential regulatory network of SWI/SNF complex and TFs (Eomesa, Pou5f3, Sox2 and Nanog) during early zebrafish development. Taken together, we reveal enriched motifs for both known and unknown factors, providing candidates of interest for future work. We note that genetic null mutants in these factors are either not publicly available, or do not create healthy maternal zygotic null animals appropriate for feasible functional studies in very early embryos, preventing a simple assessment of recruitment.

2.3.4 Correlation of Brg1 and Brm occupancy with gene expression, open chromatin and active histone marks

As SWI/SNF slides and ejects nucleosomes to facilitate the binding of transcription machinery, we investigated whether Brg1- and Brm-occupied genes have increased expression. To test, we ranked all annotated RefSeq genes by expression level (FPKM) at 5.3hpf, and compared to occupancy of Brg1, Brm and RNA Polymerase II (Pol II), centered at the transcription start site (TSS). Notably, highly-transcribed genes have higher occupancy of Brg1, Brm and Pol II at promoters. We then focused on genes occupied by Brg1 or Brm at MBT or post-MBT. In keeping, genes bound by Brg1 or Brm have higher expression compared to the genes not bound by Brg1 or Brm, or to the genome average (Figure 2.4B, C).

To reveal whether Brg1 and Brm binding is correlated with increased chromatin accessibility, we performed ATAC-seq in embryos at pre-MBT, MBT and post-MBT stages. As expected, highly expressed genes displayed more accessible and broad promoters (Figure 2.4A, blue heatmap). In keeping, we found Brg1- and Brm-occupied regions have higher ATAC-seq signals (chromatin opening) compared to the neighboring areas (Figure 2.4D). These maps also reveal clear phasing of nucleosomes into the coding region of the genes, a feature revisited later.

SWI/SNF-bound regions are correlated with histone modifications in yeast, metazoans and mammals. To examine zebrafish early embryos, we processed public ChIP-seq data of H3K27ac, H3K4me1, H3K4me3 and H3K27me3, and profiled their occupancy, centered at the TSS (Figure 2.4A, right panel) (Zhang et al., 2014; Bogdanović et al., 2012). We observed correlation of highly transcribed genes with multiple active marks (H3K27ac, H3K4me1 and H3K4me3) but not a repressive mark (H3K27me3) around their promoter regions. To examine all regions (not just

promoters) we then profiled the occupancy of histone marks around all Brg1- and Brm-occupied regions at all three stages. Intriguingly, we discovered that active marks, especially H3K27ac, are highly enriched at Brg1-occupied regions, but enrich to a much lesser extent at Brm-occupied regions (Figure 2.4E), explored further below. Taken together, Brg1 and Brm occupancy correlates with higher gene expression and increased promoters accessibility, and for Brg1-occupied loci also with the presence of active histone marks.

2.3.5 Enrichment of positive histone marks at Brg1-bound regions

To interrogate the difference between Brg1 and Brm and histone marks in more depth, we partitioned Brg1- or Brm-occupied regions based on the presence or absence of H3K27ac and H3K4me1 (Figure 2.5A), marks associated with enhancers. Notably, the large majority (71.2%) of Brg1-occupied regions bore both marks and 87.3% bore at least one, whereas only 40.8% of Brm-occupied regions had both marks and 44.2% had neither mark. We find Brm and Brg1 (in aggregate) present at a minority (15.2%) of active enhancers (enriched for both H3K27ac and H3K4me1) but virtually absent ($< 2\%$) at poised enhancers (enriched only for H3K4me1) (Figure 2.5B).

To further understand the differences between Brg1 and Brm, we separated the occupied regions at MBT into intergenic and genic, and further classified them into three types: 'Common', Brg1-specific and Brm-specific regions. Strikingly, both 'Common' and Brg1-specific regions were highly enriched for active marks, whether present in intergenic or genic areas (Figure 2.5C, D, E, Figure 2.S12). A similar pattern was also observed in post-MBT regions (Figure 2.S10, 13). Al-

though many Brm-specific regions reside in gene bodies, Brm-specific regions in promoters still displayed considerably lower enrichment for active marks than Brg1-specific regions (Figure 2.S11,12,13).

Regarding gene-function correlations, GO analysis showed that 'Common' genes are enriched for housekeeping functions (e.g. ribosome and protein biosynthesis) and Brm-specific genes were enriched for both housekeeping functions and development regulation (Figure 2.5F). In contrast, Brg1-specific genes were specifically enriched for transcription regulation and development regulation (Figure 2.5F). Of particular interest, 'Common' genes displayed substantially higher expression compared to Brg1- and Brm-specific genes, suggesting an additive/cooperative effect of Brg1 with Brm on gene expression (Figure 2.5G).

2.3.6 Occupancy of Brg1 and Brm at key pluripotency factors and miR-430, but not HOX loci

We now return to occupancy profiles at genes, and highlight notable occupied and nonoccupied genes. First, Brg1 and Brm occupy many key development regulators (Figure 2.6A), including the core factors linked to pluripotency (*nanog*, *pou5f3*, several *sox* genes), several Fox-family transcription factors (e.g. *foxh1*, *foxd3*), epigenetic regulators (*anp32e*, *smarca4a/brg1*) and signaling molecules (*wnt11*, *fgf3* and etc.). As elucidated in the GO analysis, Brg1 and Brm also occupied housekeeping genes, such as *actb1*, *actb2*. Most surprisingly, we also found Brg1 and Brm are highly enriched at the miR-430 locus in a stage-specific manner. Notably, robust Pol II binding at miR-430 cluster is observed at pre-MBT, which is consistent with previous work identifying miR-430 cluster as highly transcribed before zygotic genome activation. Curiously, little or no enrichment was observed

at the zebrafish hox loci at all three stages – reinforcing the observation that Brg1 and Brm are notably absent from almost all regions marked by H3K27me3 in the early embryo (Figure 2.S14). Taken together, Brg1 and Brm occupied a wide range of genes, suggesting involvement in cell homeostasis, pluripotency factors, regulators of lineage specification, and the transcription of miR-430, to degrade maternal RNA and promote the maternal-zygotic transition.

2.3.7 Precise co-occupancy of Brm and RNA Pol II at gene bodies

Perhaps the most striking observation in our datasets is the distribution of Brm over a high fraction of gene bodies, with nearly precise colocalization with RNA Pol II, a feature distinctly different from Brg1 (Figure 2.6). The generality of this observation can be observed in a meta-gene analysis, which displays high Brg1 enrichment specifically at promoter and very early coding region, whereas Brm has enrichment not only at promoters but also throughout gene body and into the downstream areas (3'UTRs and beyond; Figure 2.6B). We then classified genes into promoter genes (genes with only the promoter occupied) and body genes (genes with both promoter and body occupied). In keeping, the percentage of Brm Body genes is much higher than Brg1 Body genes at all three stages (Figure 2.6B), and in those cases where Brg1 displays occupancy in the body of a gene, it is typically restricted to $\sim$1kb in the promoter-proximal region.

As previewed above, we also observed a striking correlation between Pol II and Brm at Body genes, where these two factors almost always start and end at the same location (e.g. Figure 2.6A). This property is more generally displayed in metagene heatmaps for all the Brm Body genes and Promoter genes, which reveal

a high correlation between Brm and Pol II (Figure 2.S15A), which together suggest a role for Brm in transcription elongation. In support, Brm-occupied Body genes show significantly higher expression than either Brm Promoter genes or all genes. (assessed by FPKM at MBT and post-MBT, Figure 2.6D). In sum, these data reveal striking co-incidence of Brm and Pol II over the bodies of highly-transcribed genes, suggesting a role in transcription elongation.

2.3.8 Brm occupancy over gene bodies requires transcription elongation

Next, we investigated how Brm is recruited to gene body. In principle, Brm could travel with Pol II during transcription elongation, or alternatively, Brm could be recruited to gene body by histone marks such as H3K36me3, which is established co-transcriptionally by the KMT3a/Set2 complex, but remains after Pol II passage. To determine which is the underlying mechanism, we performed ChIP-seq of Brm, Pol II and H3K36me3 in flavopiridol-treated embryos (an inhibitor of transcription elongation; initial pilot experiments established dose and time parameters). To test, embryos were left untreated until 3.75hpf, which allowed both the 'minor wave'/pre-MBT transcription, as well as 'major wave'/general transcription for 45 min, and enabled the establishment of H3K36me3 marks on transcribed genes (Figure 2.7A). Embryos were then briefly treated with flavopiridol to inhibit the transcription elongation, which arrested the embryo at MBT (4hpf). Embryos were then processed for ChIP-seq of Brm, Pol II and H3K36me3. Here, if Brm travels with the Pol II transcription machinery, Brm occupancy should be lost at gene bodies only in drug-treated embryos. In contrast, if Brm is alternatively recruited by H3K36me3 (or an alternative transcription-dependent mark) post-transcriptionally,

Brm occupancy should be retained at gene bodies.

First, we observed clear effectiveness of our flavopuridol treatment, as Pol II converted from promoter and body occupancy (at Body genes) to promoter-specific occupancy only in drug-treated embryos. Here, our data clearly support the first hypothesis; Brm is recruited to the gene body by Pol II itself. In untreated embryos, Brm was highly correlated with Pol II and H3K36me3 (Figure 2.7B, left three heatmaps). Once elongation was inhibited, although some reduction in H3K36me3 was evident, a considerable fraction of body genes still retained moderate-high H3K36me3 marks in their gene bodies. However, neither Pol II or Brm displayed appreciable occupancy within gene bodies. Instead, both Pol II and Brm were largely restricted to the promoter, a feature clear in heatmaps of the entire dataset (Figure 2.7B, right three heatmaps), and also at individual genes (Figure 2.7C). Taken together, Brm appears to travel with RNA pol II during transcription elongation at $\sim$35% of all transcribed genes (Figure 2.S16), with a strong preference for the most highly transcribed genes.

2.3.9 Brm-occupied body genes are short, and have fewer exons

We then sought features that might explain why only a subset of genes are bound by Brm along the gene body. Interestingly, we found that Body genes in general are significantly shorter than promoter genes, and also have significantly fewer exons than promoter genes (Figure 2.7D). We showed earlier that Brm-occupied body genes have higher transcription, suggesting Brm might be required to facilitate fast Pol II progression through gene bodies, perhaps to contend with nucleosomes in front of Pol II. However, these Brm- and Pol II-occupied body

genes are not nucleosome-free, as our ATAC-seq maps show clear phasing of nucleosomes in the gene body, and lack general openness (Figure 2.4A). Taken together, we speculate that a considerable fraction of genes in early development have evolved to be short, contain fewer exons, and to utilize Brm remodeling to slide nucleosomes to facilitate their high transcription during a unique period of development that includes exceptionally short cell cycles and frequent competition of RNA polymerase with the replisome.

2.4 Discussion

Vertebrate embryos pass through a phase after fertilization that lacks transcription, during which their inherited parental/gametic genomes are reprogrammed at the chromatin level, to prepare genes for the regulated 'waves' of gene expression that occur at and after initial zygotic/embryonic transcriptional activation (Tadros & Lipshitz, 2009). This period of transcriptional inactivity can be relatively short (e.g. two cell cycles in humans) or relatively long, such as the 9-10 cell cycles for many fish and frogs. A key question is how and where transcription factors, chromatin modifiers and chromatin remodelers sculpt the landscape prior to and during the onset of embryonic transcriptional activation to properly identify and regulate enhancers and genes. Recently, multiple labs have begun to investigate this process in several organisms, including zebrafish, where genomic profiles and relationships have been examined between DNA methylation, nucleosomes and transcription (Vastenhouw et al., 2010; Lindeman et al., 2011; Lee et al., 2013; Leichsenring et al., 2013; Potok et al., 2013; Zhang et al., 2014). Here, we greatly extended on those studies by examining chromatin accessibility (ATAC-seq), pre-

MBT RNA Pol II – and for the first time in any cell type or embryo – the chromatin remodelers Brg1 and Brm together. Together, our work has provided several new insights into chromatin reprogramming and remodeling prior to and during transcriptional onset, and revealed clear differences between Brg1 and Brm occupancy in the early embryo genome – highlighted by the selective association of Brg1 with active histone marks, and the co-incidence of (and reliance on) RNA Pol II for Brm occupancy of coding regions and 3'UTRs of short highly-transcribed genes. Here, we find that Brg1 and Brm occupy certain loci during the pre-MBT phase, including *nanog* and *miR-430*; notably, virtually all genes with expression during the pre-MBT stage are bound by Brg1, Brm, or both. In fact, the *miR-430* locus is the most highly transcribed locus in the early 'minor' wave of transcription during pre-MBT (Heyn et al., 2014). Locations of Brg1 and Brm binding increase greatly at MBT, and include particular developmental transcription/chromatin factors and housekeeping genes. Regarding gene categories, bound genes include pluripotency factors (*nanog, pou5f3* (the zebrafish POU5F1/OCT4 ortholog)), several Fox-family transcription factors (e.g. *foxh1, foxd3*), particular epigenetic regulators (*anp32e, smarca4a/brg1*). Furthermore, Brg1 and Brm occupy a set of genes encoding particular signaling molecules, as well as many housekeeping genes. Again, virtually all occupied genes are active – either in the 'minor' wave or 'major' wave of transcription. Furthermore, genes co-occupied by Brg1 and Brm display a significantly higher level of transcription than do nonoccupied active genes. In contrast, Brg1 and Brm were generally absent from H3K27me3-marked regions in the early embryo, including zebrafish *Hox* loci and the promoters of many Tbx- and Gata-family transcription factors. Thus, Brg1 and Brm function (by localization)

appears focused on short-term preparation and execution of activation, and not long-term poising or repression of developmental genes. Interestingly, Brg1 and Brm differ greatly in their correlations with active histone marks. Brg1-specific regions, as well as regions co-occupied by both Brg1 and Brm, were highly enriched for active marks. We also observed that more than 80% of intergenic Brg1-bound regions bear at least one of two histone marks associated with enhancers, H3K4me1 and H3K27ac. Thus, our observations are consistent with prior work showing Brg1 occupancy at enhancers (Bultman et al., 2005; Rada-Iglesias et al., 2011). In counter distinction, Brm-specific regions displayed considerably lower enrichment of active marks than Brg1-specific regions. Our results are consistent with Brg1 displaying more interactions and/or cooperativity with histone modifiers that promote enhancer marking and/or transcription initiation than Brm. The most striking result in our datasets is the localization of Brm over a high fraction of gene bodies, with nearly precise co-localization with RNA Pol II, a feature distinctly different from Brg1. Our functional experiments involving inhibition of Pol II elongation strongly suggest that gene body occupancy of Brm is due to association of Brm with RNA Pol II, or with Pol II-associated elongation factors, on coding regions rather than post-transcriptional recruitment through H3K36me3 deposition. We provide two possible reasons why Brm might travel with Pol II over gene bodies. First, Brm may play a role in splicing, as prior work in a human cell line showed that overexpression of Brm increased the inclusion of variant exons. Furthermore, Brm physically interacted with splicing machinery, and was shown by ChIP-qPCR to bind the coding region of a gene, CD44 (Batsché et al., 2006). Importantly, zebrafish Brm shares high homology with the small region

at the C-terminus of human Brm, which promotes splicing (Batsché et al., 2006). Second, Brm likely has roles other than regulating splicing, as a notable minority ($\sim$6%) of Brm-containing Body genes lack introns. Here, we speculate that Brm may also be required in embryos to facilitate rapid transcription elongation through chromatin. This interpretation is based partly on RNA polymerase transcription rates; although direct measurements have not been made in zebrafish embryos, those measured in other vertebrates generally range between 1-2kb/min (Shermoen & O'Farrell, 1991). Importantly, early embryos exhibit complete cell cycles of only $\sim$15 min at pre-MBT, and $\sim$20-30 min at MBT, which puts a large fraction of transcribed genes at risk for collision events between RNA and DNA polymerase during replication. Although transcription complexes can survive DNA polymerase passage, minimizing collisions by facilitating transcription and splicing efficiency would provide a major fitness advantage. In support, Brm-occupied 'Body' genes display a strong and significant bias for short, highly-transcribed genes with fewer exons. Indeed, others have previously noted the correlation between short gene size and transcription in zebrafish and Drosophila at these developmental phases (De Renzis et al., 2007; Heyn et al., 2014), also citing the intuitive competition between transcription and replication – and here we provide a candidate molecular mechanism to help minimize this competition. Taken together, we suggest the following model for specialized Brg1 and Brm function in early embryos. Brg1 occupancy is largely restricted to enhancers and promoter regions, and is almost always co-localized with active histone modifications, perhaps due to its ability to interact with histone modifiers, or its preferred interaction with transcription factors that do so. Brm is likewise recruited to certain enhancers and promoters, but

is only modestly correlated with active histone modifications compared to Brg1. Instead, Brm is uniquely specialized to travel with RNA Pol II through to the TTS. Here, Brm may facilitate fast Pol II progression through gene bodies, possibly by helping Pol II to contend with nucleosomes and avoid pausing, but does not itself create nucleosome-free regions within coding regions (Figure. 2.4A). Our work raises several new areas of investigation regarding Brg1 and Brm specialization and transcription – where the zebrafish offers unique advantages. First, it will be of high interest to determine how Brm is specialized to interact with RNA Pol II itself (or one of the many known Pol II-associated elongation factors) in early embryos, and to further explore whether and how Brm ATPase activity might help contend with nucleosomes in coding regions. Second, as interaction of the splicing machinery and mammalian Brm has only been tested at a few select genes, the zebrafish system provides the opportunity to explore this function genome-wide and in early embryogenesis. Third, recent advances in epitope tagging genes in the zebrafish genome may allow a direct examination of transcription factor recruitment of Brg1 and Brm complexes, to better understand how focal recruitment of chromatin remodelers might help sculpt the chromatin landscape during this unique phase of development, prior to transcription, to prepare the genome for transcription onset.

2.5 Methods

2.5.1 Zebrafish and sample collection

Tubingen zebrafish lines were maintained and raised under standard conditions. Wildtype embryos were collected after 10 min of mating to ensure the

synchrony of embryo development, raised in embryo water at 28°C and staged as previously described (Kimmel et al., 1995).

2.5.2 Generation of polyclonal antibodies to Brg1 and Brm

For Brg1, 35 amino acids from the N-terminal (S155-P189) and 40 amino acids from the C-terminal (H1385-K1424) were chosen as the antigen. For Brm, 37 amino acids from the N-terminal (S255-P291) and 40 amino acids from the C-terminal (Q1529-E1568) were chosen as the antigen. The corresponding DNA sequence were synthesized, PCR amplified and cloned through BamHI and NotI sites into pETDuet-1 vector (Novogen, #71146-3), containing a N-terminal tag of 10 histidines. The constructs were transformed into BL21-CodonPlus Competent Cells, and recombinant proteins were expressed and purified via Ni/NTA affinity under denaturing conditions. The purified recombinant proteins were sent to Open Systems, Thermo Fisher Scientific using 2 rabbits on a 90-day protocol for polyclonal antibody generation.

Day90 serum was affinity purified using NHS-activated Sepharose^{TM} 4 Fast Flow (GE, 17-0906-01). Peptides used as antigen for antibody generation were equilibrated in the coupling buffer (0.2M NaHCO3, 0.5M NaCl pH8.3) using PD-10 Desalting columns (GE, cat#17-0851-01). The peptides were then coupled to the NHS-activated Sepharose beads according to the manufacturer's procedure. Day90 serum was diluted 1:1 in cold PBS, and incubated with the peptide-coupled beads at 4°C overnight. Next day, after extensive washing with cold PBS, the antibodies were eluted by elution buffer (100mM Glycine, pH2.5), and were immediately neutralized by 1M Tris [pH8.8], followed by dialysis against PBS. Purified

antibodies were stored at -80°C.

2.5.3 Antibodies and Western blotting

Antibodies used were histone H3 (Abcam ab1791), H3K36me3 (Abcam ab9050), RNA pol II (8WG16/Covance MMS-126R), phosphoserine 5 version of RNA pol II (H14/Abcam ab24759), phosphoserine 2 version of RNA pol II (H5/Abcam ab24758) and a-tubulin (Abcam 4074).

Western blotting was performed according to standard procedures, and protein sample preparation is described as below. In Figure 2.S1C, embryos around 6hpf were collected, dechorinated, and deyolked as previously described (Link et al., 2006). The cell pellet then was resuspended in lysis buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, 1% Triton, 0.1% SDS) and lysed on ice for 10 min. In Figure 2.S2B, 30 embryos at each stage were dechorinated and directly boiled at 95$^{c}ircC$ for 10 min in 2X sample buffer.

2.5.4 Chromatin immunoprecipitation

Embryos at pre-MBT (2.5hpf), MBT (4hpf) and post-MBT (5.3hpf) were carefully staged, dechorinated by Pronase (Sigma-Aldrich 11459643001), and fixed in 1% formaldehyde for 10 min at Room Temperature (RT). Formaldehyde was quenched by adding glycine to a final concentration of 0.125M, followed by centrifugation at 500g 5 min 4°C. The embryo pellet was resuspended in ice-cold PBS, and centrifuged again at 500g 5 min 4°C. The washing was repeated twice, and the embryos were snap frozen in liquid nitrogen, until enough embryos were collected.

For chromatin extraction, embryo pellets were resuspended in cell lysis buffer (10mM Tris-HCl [pH 8.1], 10mM NaCl, 0.5% NP40) and lysed for 10 min, rotating

at 4°C. Nuclei were collected by centrifugation at 1300g 5 min, 4°C, and washed twice by rotating at RT 5 min in nuclei wash buffer (50mM Tris-HCl [pH 8.0], 100mM NaCl, 10mM EDTA). The clean nulcei pellet were then resuspended in nuclei lysis buffer (50mM Tris-HCl [pH 8.1], 10mM EDTA, 1%SDS) and lysed for 10 min on ice. Samples were diluted 10 times in IP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2mM EDTA, 16.7mMTris-HCl [pH 8.1], 167mM NaCl) and sonicated with a Branson sonicator to obtain fragments around 300 bp. The fragmented chromatin were precleared by incubation with Dynabeads 1hr at 4°C (Dynabeads Sheep Anti-Rabbit IgG (11204D) for Brg1, Brm and H3K36me3, Dynabeads Pan Mouse IgG (11042) for Pan-PolII). Next, the chromatin was immunoprecipitated using the desired antibody through overnight rotation at 4°C, followed by pull-down with dynabeads during a 4-6hr rotation at 4°C.

Bound complexes were extensively washed six times with RIPA buffer (10mM Tris-HCl [pH 7.5], 140mM NaCl, 1mM EDTA, 0.5mM EGTA, 1% (vol/vol) Triton X-100, 0.1% (wt/vol) SDS, 0.1% (wt/vol) Na-deoxycholate), two times with WB LiCl (10mM Tris-HCl [pH 8.0], 1mM EDTA, 250mM LiCl, 0.5%NP-40), and one time with TE (10mM Tris-HCl [pH 8.0], 1mM EDTA). Then, the complexes were eluted from the beads by elution buffer (10mM Tris-Cl [pH 8.0], 5mM EDTA, 300mM NaCl, 0.1% SDS). The eluate was RNase treated (Ambion, AM2286) for 2hr at thermomixer (37°C 400rpm 2hr), followed by 2hr proteinase K treatment at 55°C, reverse crosslinked overnight at 65°C, and then purified by SPRI beads (Beckman Coulter, A63881).

2.5.5 ATAC-seq

ATAC-seq was performed essentially as previously described (Buenrostro et al., 2013). Briefly, embryos at pre-MBT (2.5hpf), MBT (4hpf) and post-MBT (5.3hpf) were carefully staged, and dechorinated. The embryos were washed in cold PBS, and then resuspended in lysis buffer (10mM Tris-HCl [pH 7.4], 10mM NaCl, 3mM MgCl2, 0.1% IGEPAL CA-630) on ice for 10 min. During the lysis, a 20 Gauge syringe was used three times to break the chorions and cells, followed by table top spin for 10 seconds. The supernatant was then centrifuged at 1300g 5 min 4°C and nuclei was resuspend in 2X TD buffer, and incubated with Tn5 enzyme for 30 min at 55°C (Nextera DNA Library Preparation Kit, FC-121-1031). The reaction was terminated by adding SDS to a final concentration 1%. Then, the sample was purified by Qiagen minElute column and PCR amplified by NEBNext High-Fidelity 2X PCR Master Mix (NEB, M0541L). qPCR was utilized to determine the optimal PCR cycles to prevent over-amplification.

2.5.6 Transcription elongation inhibition

Embryos were dechorinated immediately after fertilization, and closely monitored for the development stages. Flavopiridol (Sigma-Aldrich F3055) were added to the embryo water to a final concentration of 10uM to inhibit transcription elongation at 3.75hpf, which allowed 45 min genome transcription.

2.5.7 Library preparation and sequencing

Libraries of ChIP-seq samples were prepared using the NEBNext ChIP-Seq Library Prep Master Mix (NEB, E6240L). Eight ng of DNA was used and followed the manufacturer's procedure. Samples were then amplified by NEBNext High-

Fidelity 2X PCR Master Mix (NEB, M0541L) and sequenced single-end 50nt on an Illumina HiSeq 2000. For ATAC-seq samples, 101nt paired-end sequencing were performed.

2.5.8 ChIP-seq data analysis

Fastq files from ChIP-Seq were uniquely aligned to zv10 zebrafish genome assembly using Novoalign (Novocraft, Inc) with the following parameters: `-r None -Q 13 -k -o SAM`, and converted to BAM files using Samtools. BAM files from replicates were merged for downstream analysis. Peaking calling, relative to paired input, was performed using MACS2 using the following parameters: `-f BAM -g 1.4e9 -B --fix-bimodal --extsize 250 -q 0.01` (Zhang et al., 2008). Same alignment and peak calling were performed to reanalyze the previously published datasets, including Nanog ChIP-seq data (GSE34683), Eomesa ChIP-seq data (GSE51894), Pou5f3 and Sox2 ChIP-seq data (GSE39780), H3K4me3, H3K27me3 and H3K36me3 ChIP-seq data (GSE44269), H3K4me1 and H3K27ac (GSE32483).

Peak distributions were defined using RefSeq annotation (downloaded from `http://www.genome.ucsc.edu` on 15 September 2015). Customized R scripts were used to associate peaks with the closest genes. Briefly, if a peak was within $-$5kb/+5kb of a gene, it was grouped as a genic peak. If one peak could be assigned to multiple genes, only the closest gene was used. Any peaks that were not grouped as genic peaks were grouped as intergenic peaks. Distances within $-$2kb/+1kb of a gene TSS were defined as promoter. Distances within 0/+2kb of a gene TTS were defined as downstream.

For heatmap, data within $-$2kb/$+$2kb of a gene TSS, or of a peak center were

collected every 50bp in 80 windows, using the program get_relative_data (`http://search.cpan.org/dist/Bio-ToolBox/`) and were visualized by Java TreeView (Saldanha, 2004). Data in each window were averaged and graphed with GraphPad Prism (GraphPad Software, Inc.) as scatterplot to represent the averaged data distribution around genes TSS or peak centers.

For metagene model analysis, each occupied gene has been divided into 20 bins, and data within −3kb upstream TSS, and +3kb downstream TTS, and the 20bins of each gene were collected using the program get_binned_data (`http://search.cpan.org/dist/Bio-ToolBox/`).

Gene ontology analysis for all Brg1 and Brm at all stages were performed with DAVID (Huang et al., 2009). Enrichment of known motifs was analyzed using RSAT (Medina-Rivera et al., 2015).

Signal tracks representing −log(qvalue) were built using MACS bdgcmp with parameter `--m qpois`, and were visualized using Integrated Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013).

2.5.9 ATAC-seq data analysis

Paired-end fastq files were uniquely aligned to zv10 zebrafish genome assembly using Novoalign (Novocraft, Inc) with the following parameters: `-r None -Q 13 -k -o SAM -a CTGTCTCTTATACACATCT`, and converted to BAM files using Samtools. BAM files from replicates were merged for downstream analysis. Total ATAC-seq data are divided into nucleosome-free reads and nucleosome signal according to (Buenrostro et al., 2013). The coverage tracks of nucleosome-free reads were generated using the program bam2wig (`http://search.cpan.org/`

`dist/Bio-ToolBox/`) with the following parameters: `--position span --pe --rpm --bw`.

2.6 Contribution and acknowledgement

This work is under preparation for submission to *Genes & Development*. Brad Cairns and I have worked together for the writing. I have performed all the bioinformatic analysis and most of the ChIP-seq experiments. Yixuan Guo has helped me with the ChIP-seq in drug-treated embryos. Candice Wike has helped with ATAC-seq experiments. Christian Pflueger has prepared the RNA-seq data. We want to thank Tim Parnell and Tim Mosbruger for bioinformatics assistance, and Brian Dalley for sequencing expertise.

2.7 References

Batsché, E., Yaniv, M., Muchardt, C. (2006). The human swi/snf subunit brm is a regulator of alternative splicing. Nature Structural & Molecular Biology *13*, 22–29.

Bogdanović, O., Fernandez-Miñán, A., Tena, J.J., de la Calle-Mustienes, E., Hidalgo, C., van Kruysbergen, I., van Heeringen, S.J., Veenstra, G.J.C., Gómez-Skarmeta, J.L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Research *22*, 2043–2053.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. Nature methods *10*, 1213–1218.

Bultman, S.J., Gebuhr, T.C., Magnuson, T. (2005). A brg1 mutation that uncouples atpase activity from chromatin remodeling reveals an essential role for swi/snf-related complexes in β-globin expression and erythroid development. Genes & Development *19*, 2849–2861.

Bultman, S.J., Gebuhr, T.C., Pan, H., Svoboda, P., Schultz, R.M., Magnuson, T. (2006). Maternal brg1 regulates zygotic genome activation in the mouse. Genes & Development *20*, 1744–1754.

Bultman, S., Gebuhr, T., Yee, D., La Mantia, C., Nicholson, J., Gilliam, A., Randazzo, F., Metzger, D., Chambon, P., Crabtree, G., et al. (2000). A brg1 null

mutation in the mouse reveals functional differences among mammalian swi/snf complexes. Molecular Cell *6*, 1287–1295.

Clapier, C.R., Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annual Review of Biochemistry *78*, 273–304.

De Renzis, S., Elemento, O., Tavazoie, S., Wieschaus, E.F. (2007). Unmasking activation of the zygotic genome using chromosomal deletions in the drosophila embryo. PLoS Biol *5*, e117.

Eroglu, B., Wang, G., Tu, N., Sun, X., Mivechi, N.F. (2006). Critical role of brg1 member of the swi/snf chromatin remodeling complex during neurogenesis and neural crest induction in zebrafish. Developmental Dynamics *235*, 2722–2735.

Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A.T., Neugebauer, K.M. (2014). The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Reports *6*, 285–292.

Ho, L., Jothi, R., Ronan, J.L., Cui, K., Zhao, K., Crabtree, G.R. (2009). An embryonic stem cell chromatin remodeling complex, esbaf, is an essential component of the core pluripotency transcriptional network. Proceedings of the National Academy of Sciences *106*, 5187–5191.

Hodges, C., Kirkland, J.G., Crabtree, G.R. (2016). The many roles of baf (mswi/snf) and pbaf complexes in cancer. Cold Spring Harbor Perspectives in Medicine *6*, a026930.

Hota, S.K., Bruneau, B.G. (2016). Atp-dependent chromatin remodeling during mammalian development. Development *143*, 2882–2897.

Huang, D.W., Sherman, B.T., Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature Protocols *4*, 44–57.

Jaenisch, R., Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567–582.

Kadam, S., Emerson, B.M. (2003). Transcriptional specificity of human swi/snf brg1 and brm chromatin remodeling complexes. Molecular Cell *11*, 377–389.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Developmental Dynamics *203*, 253–310.

Lee, M.T., Bonneau, A.R., Giraldez, A.J. (2014). Zygotic genome activation during the maternal-to-zygotic transition. Annual Review of Cell and Developmental Biology *30*, 581.

Lee, M.T., Bonneau, A.R., Takacs, C.M., Bazzini, A.A., DiVito, K.R., Fleming, E.S., Giraldez, A.J. (2013). Nanog, pou5f1 and soxb1 activate zygotic gene expression during the maternal-to-zygotic transition. Nature *503*, 360–364.

LeGouy, E., Thompson, E.M., Muchardt, C., Renard, J.P. (1998). Differential preimplantation regulation of two mouse homologues of the yeast swi 2 protein. Developmental Dynamics *212*, 38–48.

Leichsenring, M., Maes, J., Mössner, R., Driever, W., Onichtchouk, D. (2013). Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science *341*, 1005–1009.

Lessard, J.A., Crabtree, G.R. (2010). Chromatin regulatory mechanisms in pluripotency. Annual Review of Cell and Developmental Biology *26*, 503.

Lindeman, L.C., Andersen, I.S., Reiner, A.H., Li, N., Aanes, H., Østrup, O., Winata, C., Mathavan, S., Müller, F., Aleström, P., et al. (2011). Prepatterning of developmental gene expression by modified histones before zygotic genome activation. Developmental Cell *21*, 993–1004.

Link, V., Shevchenko, A., Heisenberg, C.P. (2006). Proteomics of early zebrafish embryos. BMC Developmental Biology *6*, 1.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., et al. (2015). Rsat 2015: regulatory sequence analysis tools. Nucleic Acids Research , gkv362.

Nelson, A.C., Cutty, S.J., Niini, M., Stemple, D.L., Flicek, P., Houart, C., Bruce, A.E., Wardle, F.C. (2014). Global identification of smad2 and eomesodermin targets in zebrafish identifies a conserved transcriptional network in mesendoderm and a novel role for eomesodermin in repression of ectodermal gene expression. BMC Biology *12*, 1.

Niwa, H. (2007). How is pluripotency determined and maintained? Development *134*, 635–646.

Potok, M.E., Nix, D.A., Parnell, T.J., Cairns, B.R. (2013). Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell *153*, 759–772.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Reyes, J., Barra, J., Muchardt, C., Camus, A., Babinet, C., Yaniv, M. (1998). Altered control of cellular proliferation in the absence of mammalian brahma (snf2α). The EMBO Journal *17*, 6979–6991.

Russ, A.P., Wattler, S., Colledge, W.H., Aparicio, S.A., Carlton, M.B., Pearce, J.J., Barton, S.C., Surani, M.A., Ryan, K., Nehls, M.C., et al. (2000). Eomesodermin is required for mouse trophoblast development and mesoderm formation. Nature *404*, 95–99.

Saldanha, A.J. (2004). Java treeviewextensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Shermoen, A.W., O'Farrell, P.H. (1991). Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. Cell *67*, 303–310.

Tadros, W., Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. Development *136*, 3033–3042.

Thompson, K.W., Marquez, S.B., Lu, L., Reisman, D. (2015). Induction of functional brm protein from brm knockout mice. Oncoscience *2*, 349.

Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P. (2013). Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in Bioinformatics *14*, 178–192.

Vastenhouw, N.L., Zhang, Y., Woods, I.G., Imam, F., Regev, A., Liu, X.S., Rinn, J., Schier, A.F. (2010). Chromatin signature of embryonic pluripotency is established during genome activation. Nature *464*, 922–926.

Xu, C., Fan, Z.P., Müller, P., Fogley, R., DiBiase, A., Trompouki, E., Unternaehrer, J., Xiong, F., Torregroza, I., Evans, T., et al. (2012). Nanog-like regulates endoderm formation through the mxtx2-nodal pathway. Developmental Cell *22*, 625–638.

Zernicka-Goetz, M., Morris, S.A., Bruce, A.W. (2009). Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo. Nature Reviews Genetics *10*, 467–477.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of chip-seq (macs). Genome Biology *9*, 1.

Zhang, Y., Vastenhouw, N.L., Feng, J., Fu, K., Wang, C., Ge, Y., Pauli, A., van Hummelen, P., Schier, A.F., Liu, X.S. (2014). Canonical nucleosome organization at promoters forms during genome activation. Genome Research *24*, 260–266.

Table 2.1: Conservation of SWI/SNF complexes members between zebrafish and human

subunit	NCBI	Length	Uniprot	Length	human homology			
					ID	Length	Identity	similarity
Brg1	NP_853634.1	1,627	Q7ZSY3	1,627	NP_003063.2	1,647	86%	89%
Brm	NP_001038240.1	1,568	Q6P9P2	1,568	NP_620614.2	1,590	82%	88%
Polybromo1	XP_693735.4	1,581	NA	NA	AAG48939.1	1,602	71%	81%
Arid1aa	XP_009292406.1	2,261	A0A0R4IKJ3	2,286	AAG33967.1	2,285	62%	71%
Arid1ab	XP_009292406.1	2,261	A0A0R4IR17	2,101	NP_065783.3 (1b)	2,249	56%	67%
Arid1b	XP_009292820.1	2,121	Q5RGW7	2,121	NP_065783.3 (1b)	2,236	57%	69%
BAF155/smarcc1a	AAH47827.1	839	Q7ZUU5	839	NP_003065.3	1,105	75%	85%
BAF155/smarcc1b	NP_001082836.1	959	A0A0R4IDE0	969	NP_003065.3	1,105	65%	78%
BAF170	XP_695864.3	1,036	E7FFZ4	1,036	NP_003066.2	1,214	82%	90%
SNF5/BAF47a	NP_001007297.1	373	Q5U379	373	NP_001007469.1	376	96%	98%
SNF5/BAF47b	NP_571523.1	370	Q5XJR9	370	NP_001007469.1	376	93%	98%
BAF60a	NP_938172.1	510	Q802C8	510	NP_003067.3	515	92%	94%
BAF60b	XP_692749.2	501	E7EYW7	501	NP_001091896.1	531	78%	86%
BAF60ca	XP_005163574.1	505	E7F1T4	505	NP_001003801.1	483	81%	88%
BAF60cb	NP_001120778.1	476	B2KL77	476	NP_001003801.1	483	88%	92%
BAF53a	NP_775347.1	429	Q8JGS6	429	NP_004292.1	429	88%	94%
BAF53b	XP_005170633.1	426	F8W3Z2	426	NP_057272.1	426	86%	92%
BAF57	NP_958455.2	420	Q803S1	420	AAH07082.1	411	79%	85%

Table 2.2: ChIP-seq sequencing summary

		Rep1			Rep2			
	Sample	**Sequenced reads**	**Uniquely mapped reads**	**Map %**	**Sequenced reads**	**Uniquely mapped reads**	**Mapp %**	**Pearson Correlation**
	preMBT	22,619,187	16,178,109	71.52%				
Brg1	MBT	22,164,263	16,003,970	72.21%	30,170,317	22,575,308	74.82%	0.94
	postMBT	24,257,272	17,881,059	73.71%	38,612,645	28,936,176	74.94%	0.96
	preMBT	27,682,342	19,081,720	68.93%				
Brm	MBT	30,559,682	21,418,799	70.09%	28,680,393	21,103,953	73.58%	0.96
	postMBT	23,995,825	17,628,529	73.46%	39,188,263	29,255,118	74.65%	0.99
	preMBT	31,857,139	22,004,993	69.07%				
PolII	MBT	42,684,043	31,376,888	73.51%	36,325,631	25,974,153	71.50%	0.98
	postMBT	26,402,416	19,209,488	72.76%	29,459,020	21,791,312	73.97%	0.97

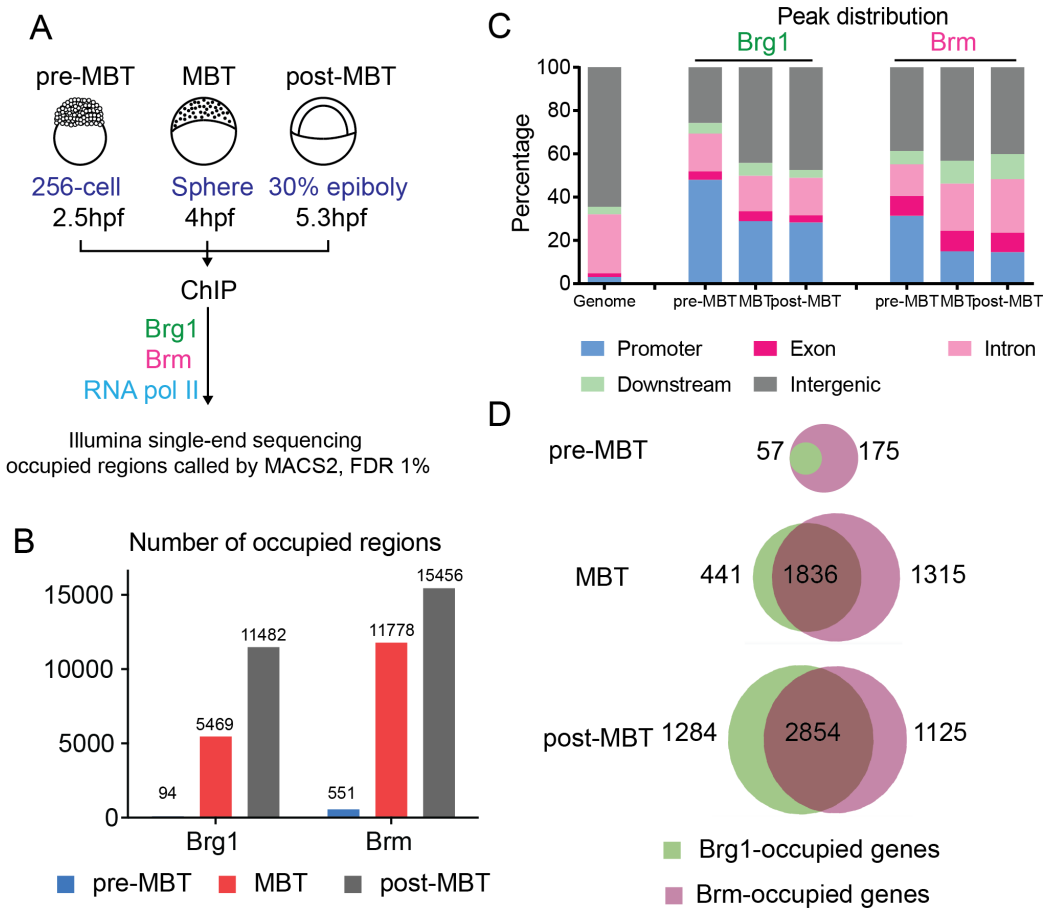
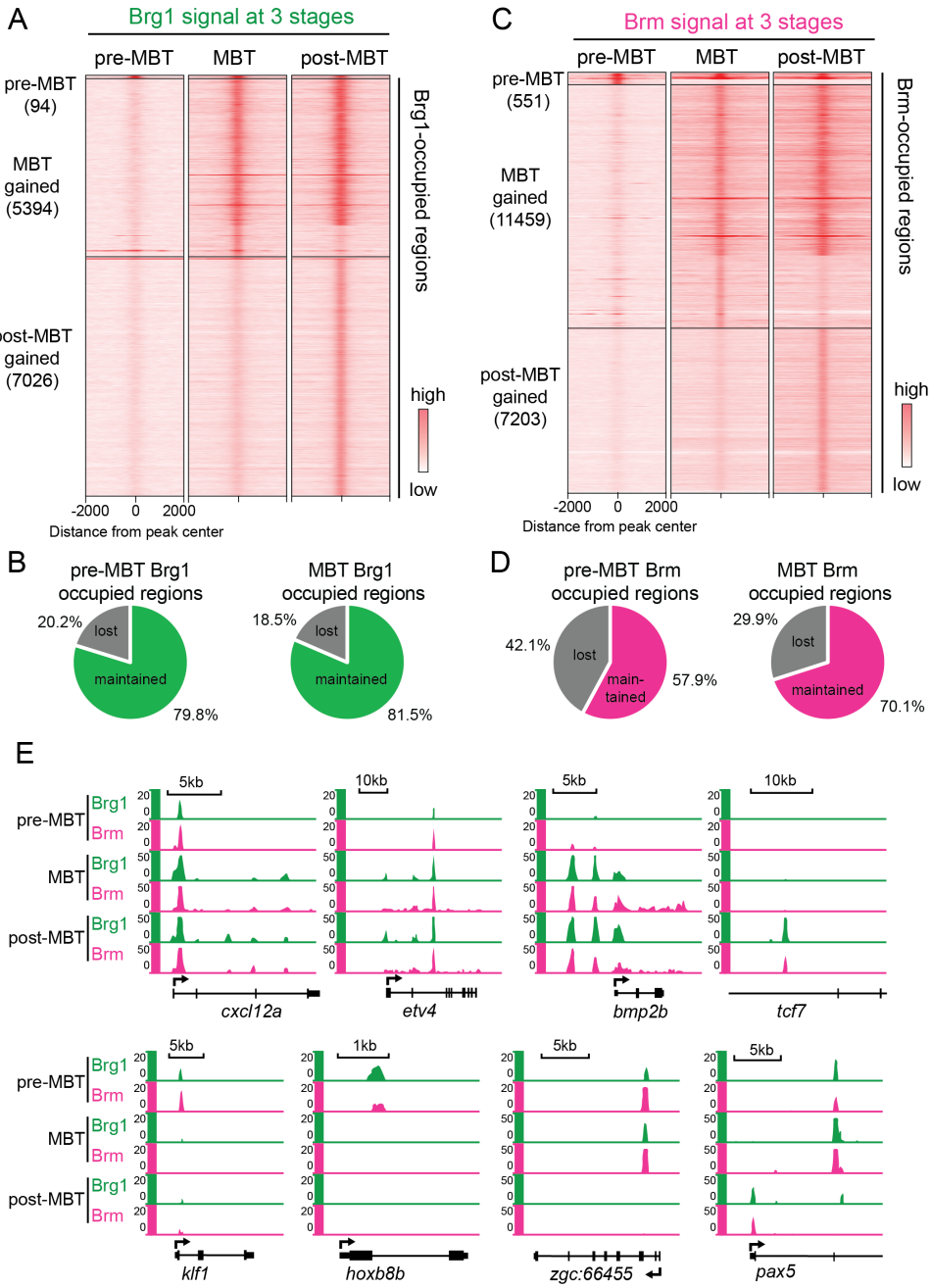


Figure 2.1: Overview of the ChIP-seq datasets. (A) Schematic of ChIP-seq approaches in embryos at pre-MBT, MBT and post-MBT stages. (B) Occupied regions of Brg1 and Brm at each stage (defined by MACS2 peak calling). (C) Distribution of Brg1 and Brm occupancy at each stage on RefSeq annotation elements. Promoters represent the regions −2kb, +1kb around the TSS. (D) Venn diagram analysis of the overlap between Brg1- and Brm-occupied genes at each stage.

Figure 2.2: Dynamic occupancy of Brg1 and Brm in early development. (A) Heatmaps of Brg1 occupancy at three stages around Brg1-occupied regions at pre-MBT, Brg1-occupied regions gained at MBT and post-MBT. Each row represents $\pm$2kb around the center of occupied regions. Each heatmap corresponds to one stage. (B) Pie charts of the percentage of maintained or lost of Brg1-occupied regions at the subsequent stage. (C) Heatmaps of Brm occupancy at three stages around Brm-occupied regions, as in (A). (D) Pie charts of the percentage of maintained or lost of Brm-occupied regions at the subsequent stage. (E) Genome browser snapshots of representative dynamic occupied regions for Brg1 and Brm in early development.


A
Brg1 signal at 3 stages
pre-MBT
MBT
post-MBT
pre-MBT (94)
MBT gained (5394)
post-MBT gained (7026)
Brg1-occupied regions
high
low
-2000 0 2000
Distance from peak center
C
Brm signal at 3 stages
pre-MBT (551)
MBT gained (11459)
post-MBT gained (7203)
Brm-occupied regions
B
pre-MBT Brg1 occupied regions
20.2%
lost
maintained
79.8%
MBT Brg1 occupied regions
18.5%
81.5%
D
pre-MBT Brm occupied regions
42.1%
main-tained
57.9%
MBT Brm occupied regions
29.9%
70.1%
E
Brg1
Brm
5kb
10kb
1kb
cxcl12a
etv4
bmp2b
tcf7
klf1
hoxb8b
zgc:66455
pax5

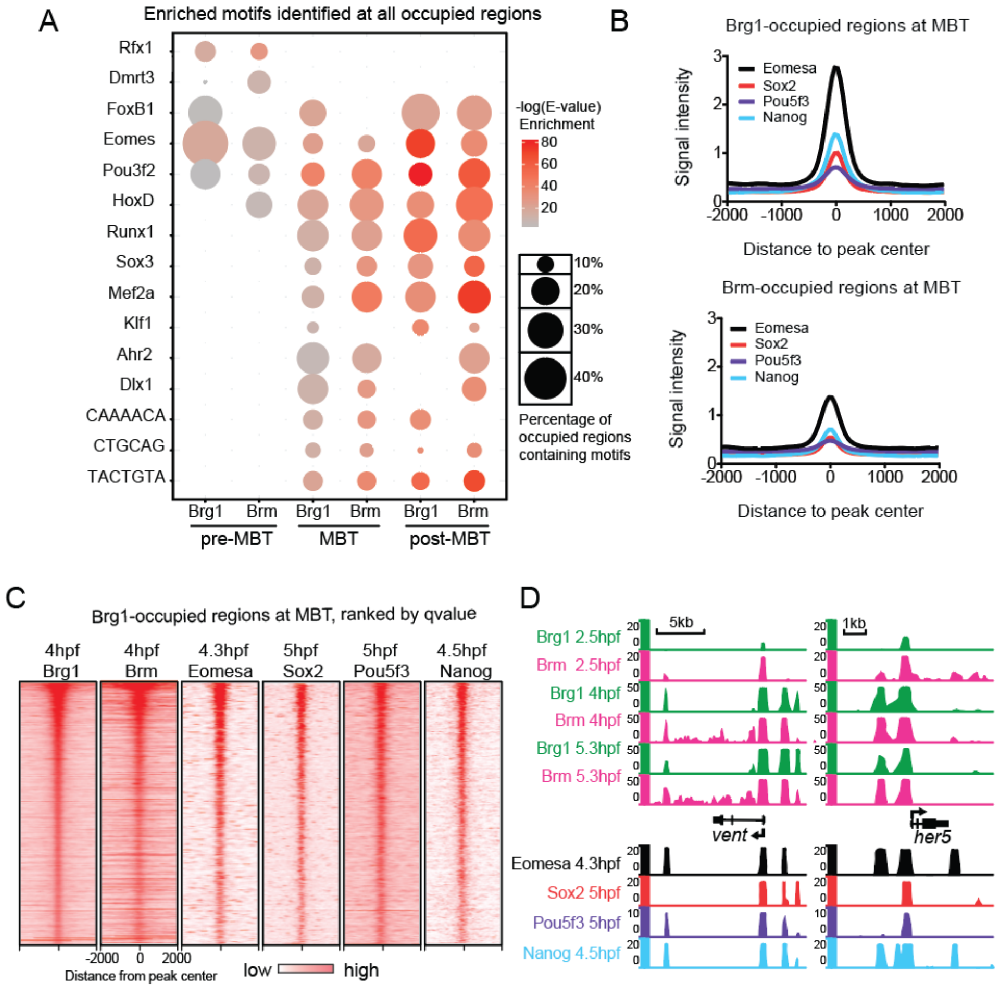


Figure 2.3: Pluripotency factors are potential Brg1 or Brm interactors. (A) Over-represented motifs identified at Brg1- and Brm-occupied regions during each stage. The color represents the enrichment P-value, and the size the percentage of occupied regions containing specific motif. (B) The mean profiles of Eomesa, Sox2, Pou5f3, and Nanog signal (ChIP-seq data) around Brg1- and Brm-occupied regions at MBT stage. (C) Heatmaps of ChIP-seq signal (from left to right Brg1, Brm, Eomesa, Sox2, Pou5f3, and Nanog) around Brg1-occupied regions at MBT, ranked by q-value. (D) Genome browser snapshots of representative co-localization of Brg1 and Brm with transcription factors (Eomesa, Sox2, Pou5f3, and Nanog).

Figure 2.4: Brg1 and Brm occupancy correlates with higher gene expression, open chromatin, and active histone marks. (A) Heatmaps of PolII, Brg1, Brm, H3K27ac, H3K4me1, H3K4me3, H3K27me3 and H3K36me3 signal (ChIP-seq) around $\pm$2kb of TSS, and ATAC-seq signal around $\pm$1kb of TSS. Each row represents a RefSeq gene TSS, ranked by RNA FPKM value at 5.3hpf (leftmost panel). (B) Boxplot of expression level of Brg1 occupied genes, nonoccupied genes at MBT and post-MBT stages, and all genes (t test: *** denotes $p < 0.001$). (C) Same as in (B) but for Brm. (D) The mean profiles of ATAC-seq signal at pre-MBT, MBT and post-MBT for Brg1- and Brm-occupied regions at each stage. (E) The mean profiles of H3K27ac, H3K4me1, H3K4me3, H3K27me3 signal (ChIP-seq) at 4.3hpf for Brg1- and Brm-occupied regions at each stage.

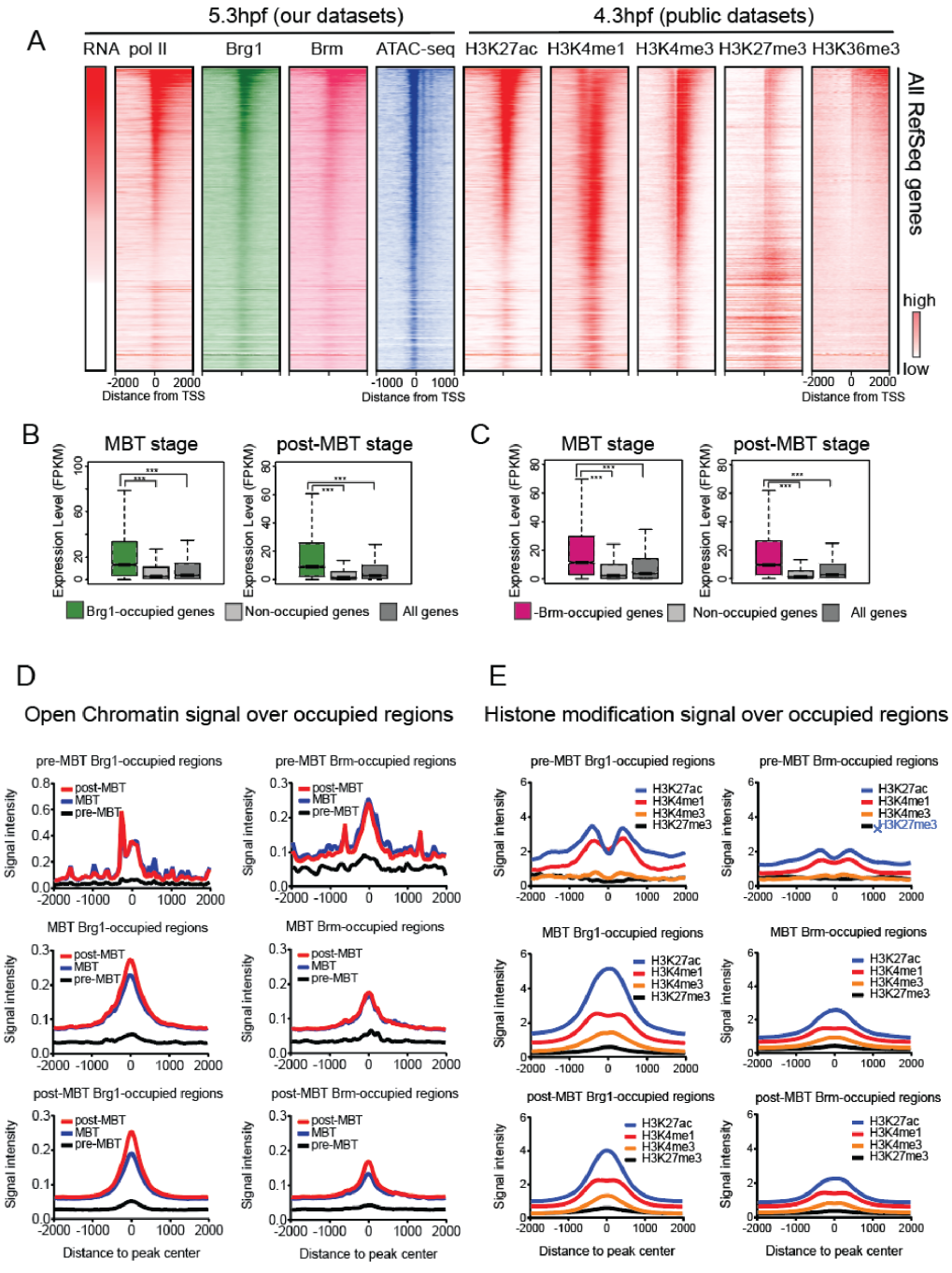

A
5.3hpf (our datasets)
4.3hpf (public datasets)
RNA pol II
Brg1
Brm
ATAC-seq
H3K27ac
H3K4me1
H3K4me3
H3K27me3
H3K36me3
All RefSeq genes
high
low
-2000 0 2000
Distance from TSS
-1000 0 1000
Distance from TSS
B
MBT stage
post-MBT stage
Expression Level (FPKM)
Brg1-occupied genes
Non-occupied genes
All genes
C
MBT stage
post-MBT stage
Expression Level (FPKM)
-Brm-occupied genes
Non-occupied genes
All genes
D
Open Chromatin signal over occupied regions
pre-MBT Brg1-occupied regions
pre-MBT Brm-occupied regions
MBT Brg1-occupied regions
MBT Brm-occupied regions
post-MBT Brg1-occupied regions
post-MBT Brm-occupied regions
post-MBT
MBT
pre-MBT
Signal intensity
Distance to peak center
E
Histone modification signal over occupied regions
H3K27ac
H3K4me1
H3K4me3
H3K27me3
Signal intensity
Distance to peak center

Figure 2.5: Brg1-bound regions are enriched for active histone marks. (A) Pie chart of the percentage of Brg1- and Brm-bound intergenic regions at MBT having different combination of H3K4me1 and H3K27ac signal. (B) Pie chart of percentage of active enhancers (marked by H3K4me1 and H3K27ac) and poised enhancer (marked only by H3K4me1) having Brg1 and Brm signal. (C) Heatmaps of Brg1, Brm, H3K27ac, H3K4me1, H3K4me3 and H3K27me3 signal (ChIP-seq) around occupied intergenic regions at MBT. Common: regions co-occupied by Brg1 and Brm; Brg1-specific: regions specifically occupied by Brg1; Brm-specific: regions specifically occupied by Brm. Each row represents $\pm$2kb around the center of occupied regions. Among each category, occupied regions are ranked by q-value. (D) The mean profiles of H3K27ac and H3K4me1 signal (ChIP-seq) for common, Brg1-specific, Brm-specific and random regions. (E) Heatmaps of signal around occupied genic regions at MBT, as in (C). (F) Boxplot of expression level of Common (genes co-occupied by Brg1 and Brm), Brg1-specific genes (genes specifically occupied by Brg1), Brm-specific genes (genes specifically occupied by Brm), and genes at MBT and post-MBT stages (t test: *** denotes $p < 0.001$).

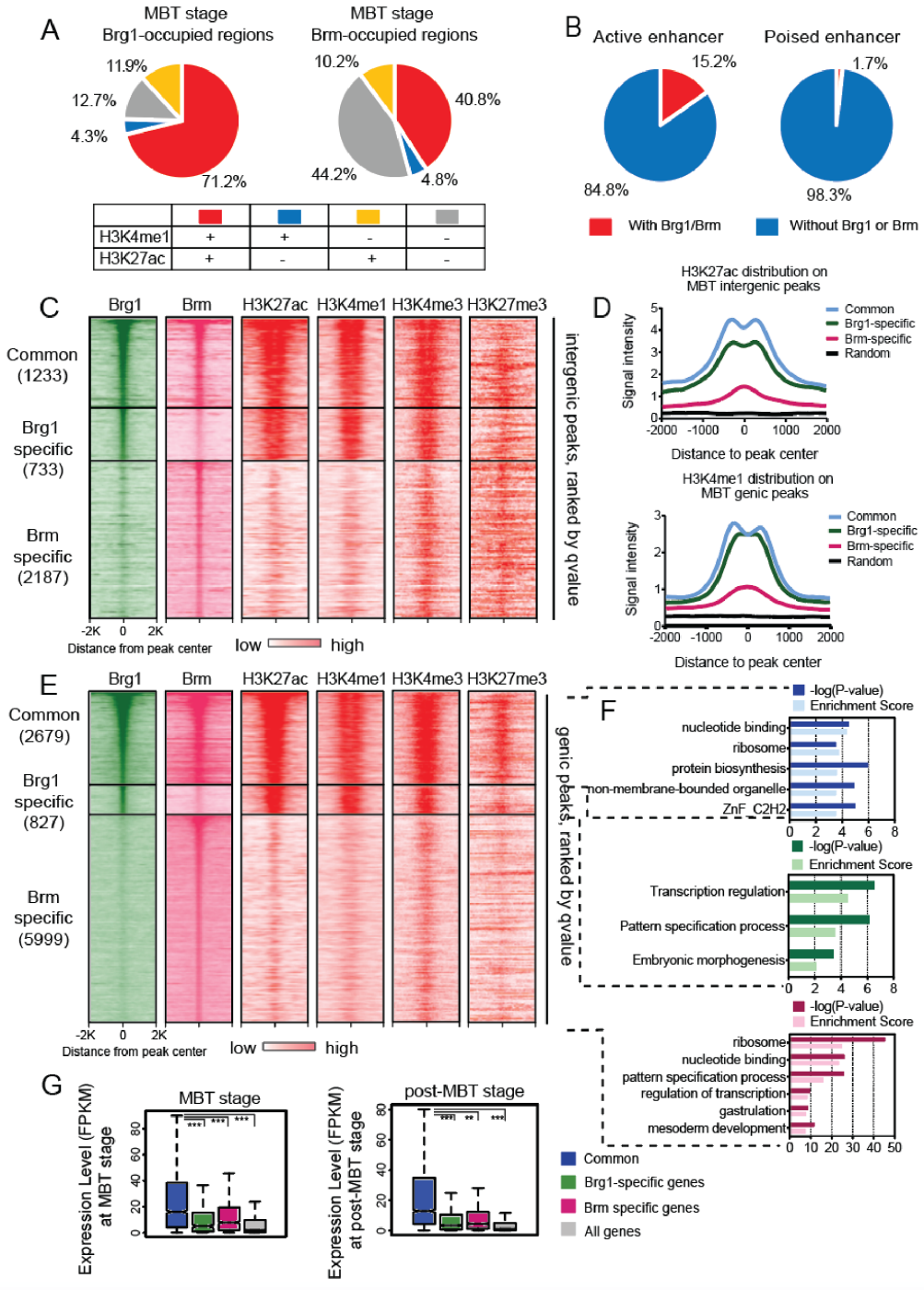

A
MBT stage
Brg1-occupied regions
11.9%
12.7%
4.3%
71.2%
MBT stage
Brm-occupied regions
10.2%
40.8%
44.2%
4.8%
H3K4me1
H3K27ac
B
Active enhancer
Poised enhancer
15.2%
1.7%
84.8%
98.3%
With Brg1/Brm
Without Brg1 or Brm
C
Brg1
Brm
H3K27ac
H3K4me1
H3K4me3
H3K27me3
Common (1233)
Brg1 specific (733)
Brm specific (2187)
intergenic peaks, ranked by qvalue
Distance from peak center
low
high
D
H3K27ac distribution on MBT intergenic peaks
Common
Brg1-specific
Brm-specific
Random
Signal intensity
Distance to peak center
H3K4me1 distribution on MBT genic peaks
E
Common (2679)
Brg1 specific (827)
Brm specific (5999)
genic peaks, ranked by qvalue
F
-log(P-value)
Enrichment Score
nucleotide binding
ribosome
protein biosynthesis
non-membrane-bounded organelle
ZnF_C2H2
Transcription regulation
Pattern specification process
Embryonic morphogenesis
ribosome
nucleotide binding
pattern specification process
regulation of transcription
gastrulation
mesoderm development
G
MBT stage
post-MBT stage
Expression Level (FPKM) at MBT stage
Expression Level (FPKM) at post-MBT stage
Common
Brg1-specific genes
Brm specific genes
All genes

Figure 2.6: Brm is highly enriched in gene bodies. (A) Genome browser snapshots of representative Brg1 and Brm occupancy on pluripotency factors (nanog and pou5f3), transcription factor (foxh1), epigenetic regulator (anp32b), microRNA (miR-430) and housekeeping genes (actb2). (B) Meta-gene model of mean profiles of Brg1 and Brm signal on their occupied genes at each stage. (C) Bar graph of percentage of Body genes (genes have Brm occupancy at promoter and gene body) and Promoter genes (genes only have Brm occupancy at promoter). (D) Boxplot of expression level of Body and Promoter genes of Brm-occupied genes at MBT and post-MBT stages (t test: *** denotes $p < 0.001$).

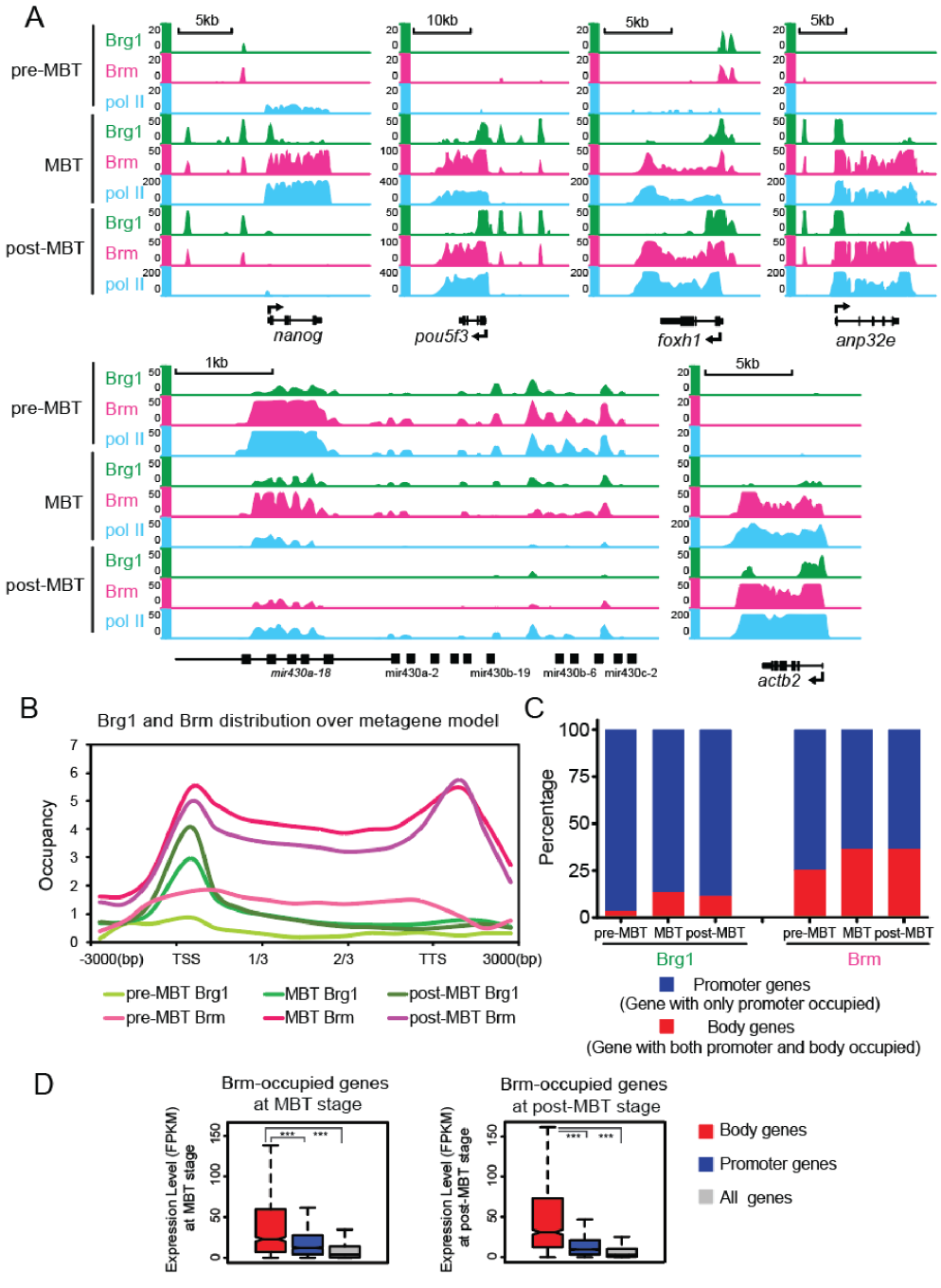
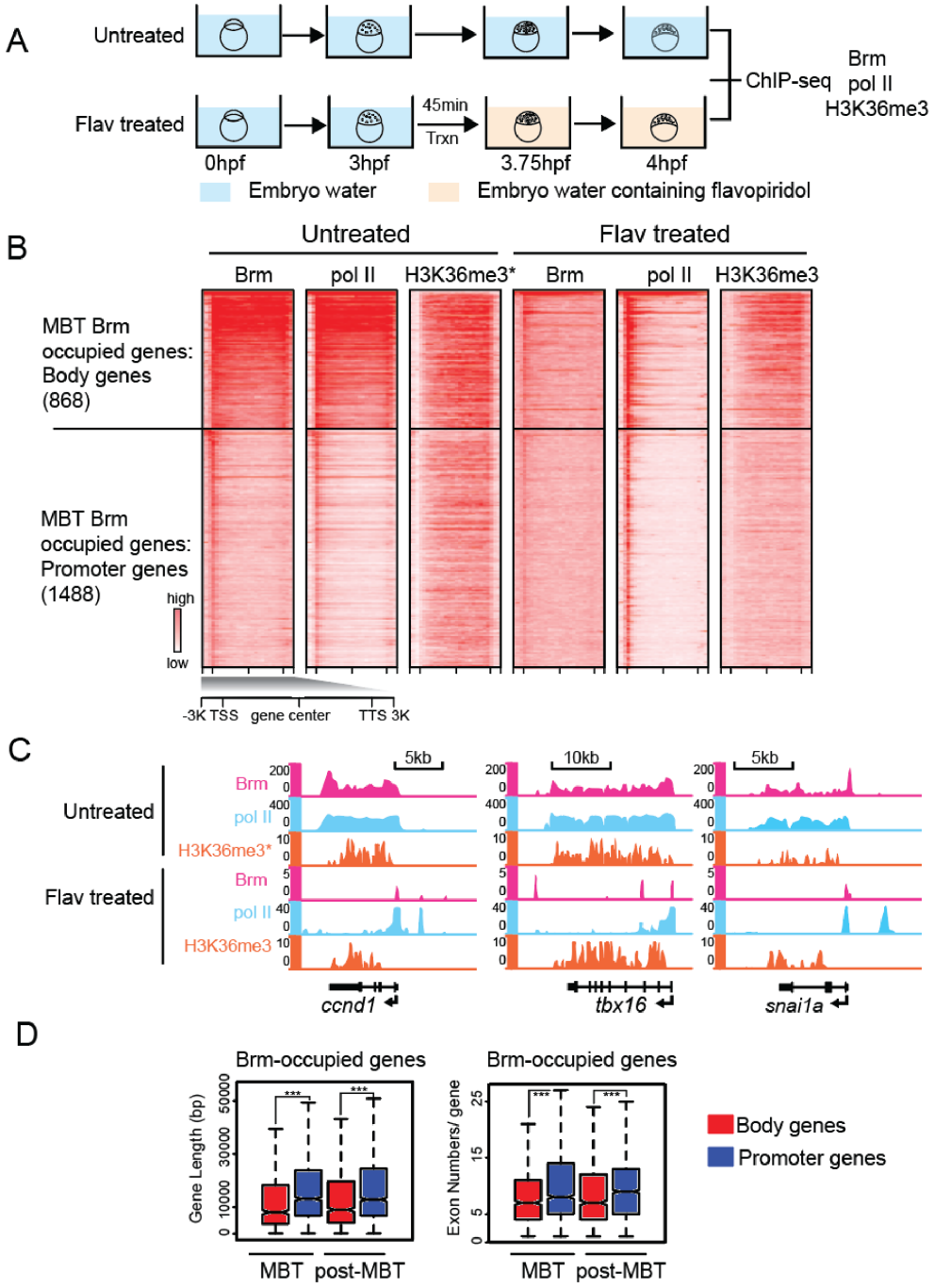

A
pre-MBT
MBT
post-MBT
Brg1
Brm
pol II
5kb
10kb
1kb
nanog
pou5f3
foxh1
anp32e
mir430a-18
mir430a-2
mir430b-19
mir430b-6
mir430c-2
actb2
B
Brg1 and Brm distribution over metagene model
Occupancy
-3000(bp)
TSS
1/3
2/3
TTS
3000(bp)
pre-MBT Brg1
MBT Brg1
post-MBT Brg1
pre-MBT Brm
MBT Brm
post-MBT Brm
C
Percentage
pre-MBT MBT post-MBT
Brg1
Brm
Promoter genes
(Gene with only promoter occupied)
Body genes
(Gene with both promoter and body occupied)
D
Brm-occupied genes at MBT stage
Brm-occupied genes at post-MBT stage
Expression Level (FPKM) at MBT stage
Expression Level (FPKM) at post-MBT stage
Body genes
Promoter genes
All genes

Figure 2.7: Enrichment of Brm over gene body is dependent on transcription elongation. (A) Schematic of elongation inhibition by flavopiridol. (B) Heatmaps of Brm, PolII, and H3K36me3 ChIP-seq signal in untreated and flavopiridol treat embryos around MBT Brm occupied Body genes and Promoter genes. Each row represents one MBT Brm-occupied gene, which includes 20 bins for entire gene, 3 bins for upstream 3kb and 3 bins for downstream 3kb. H3K36me3 ChIP-seq is performed with 4.3hpf embryos (public dataset), and all the other ChIP-seqs are performed with 4hpf embryos (our original datasets). (C) Genome browser snapshots of representative genes that have H3K36me3 but not Brm in the gene body after flavopiridol treatment. (D) Boxplot of gene length and exon numbers of Brm Body and Promoter genes at MBT and post-MBT stage.


A
Untreated
Flav treated
45min
Trxn
ChIP-seq
Brm
pol II
H3K36me3
0hpf
3hpf
3.75hpf
4hpf
Embryo water
Embryo water containing flavopiridol
B
Untreated
Flav treated
Brm
pol II
H3K36me3*
H3K36me3
MBT Brm occupied genes: Body genes (868)
MBT Brm occupied genes: Promoter genes (1488)
high
low
-3K TSS
gene center
TTS 3K
C
Untreated
Flav treated
5kb
10kb
ccnd1
tbx16
snai1a
D
Brm-occupied genes
Gene Length (bp)
Exon Numbers/ gene
MBT
post-MBT
Body genes
Promoter genes

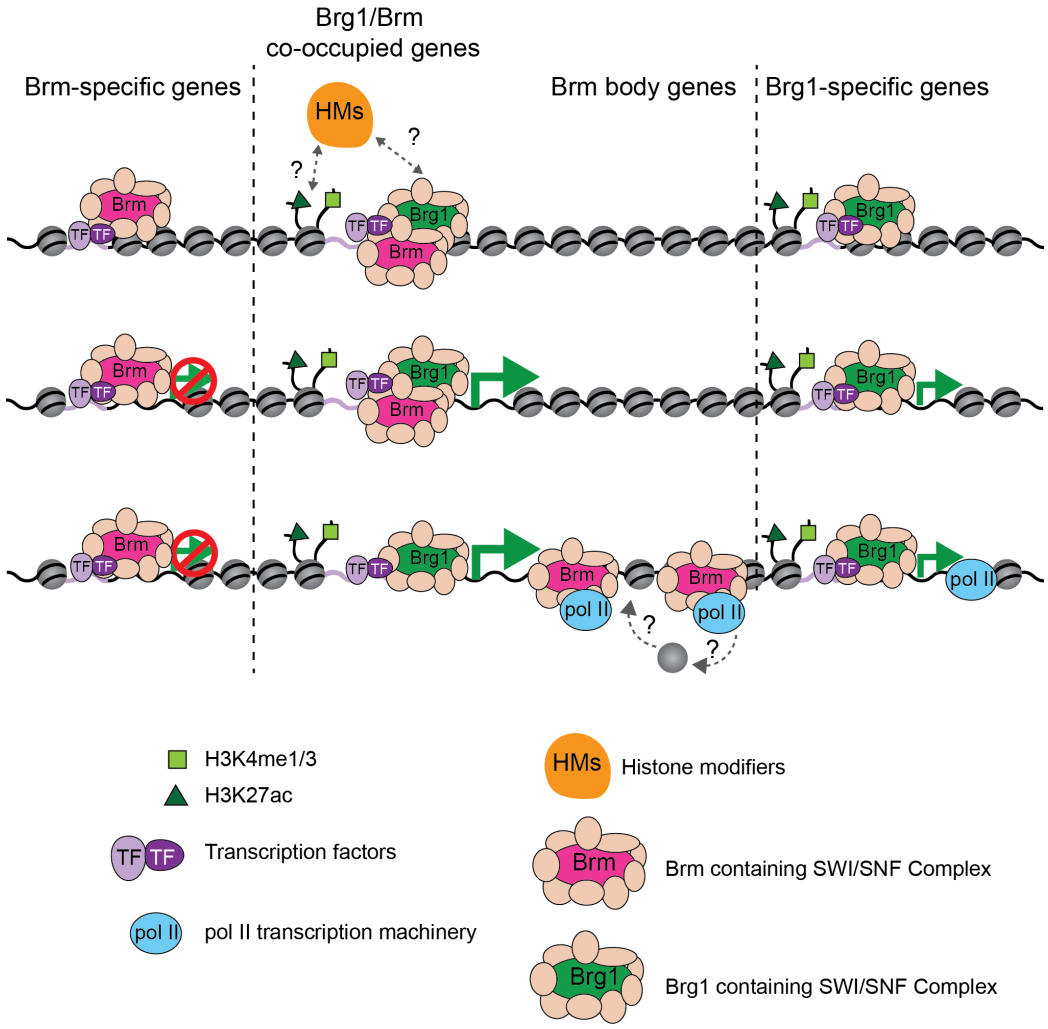
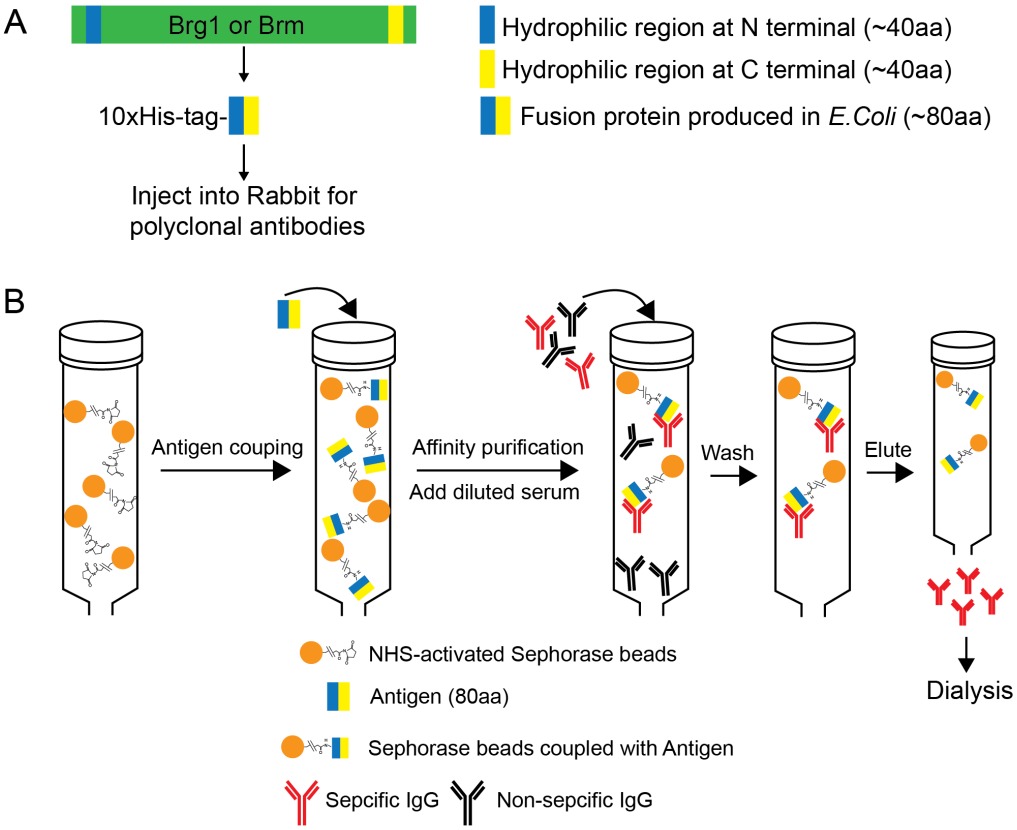
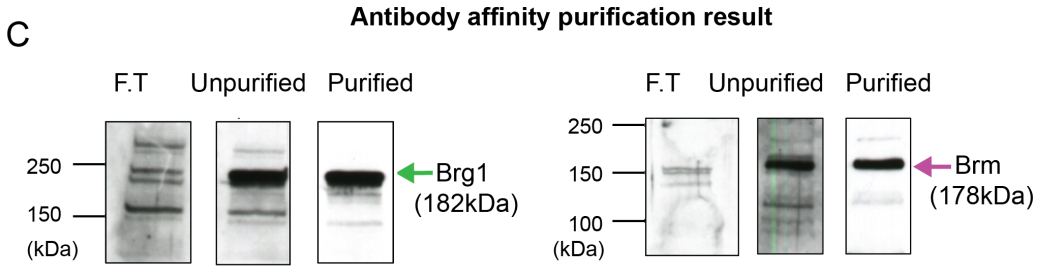


Figure 2.8: Model. (A) Brg1 and Brm overlap at most genes, but also solely/specifically occupy certain genes. (B) Regions bound by Brg1, much more so than Brm, correlate with active histone marks (H3K4me1/3 and H3K27ac), and the absence of H3K27me3. (C) Brg1-containing SWI/SNF complex is enriched at promoters, and Brm-containing SWI/SNF complex is enriched at both the promoter and also at the gene body. Genes with the highest transcription tend to have both Brg1 and Brm.

Figure 2.S1: Antibody generation and validation. (A) Schematic of strategy for polyclonal antibody generation. (B) Schematic of strategy for antibody affinity purification. (C) Immunoblotting using flow-through, unpurified antibody and purified antibody against Brg1 in 6hpf whole embryo lysate. (D) Immunoblotting against Brm, same as in (C).





F.T: Flow through
Protein sample: 6hpf whole embryo lysate

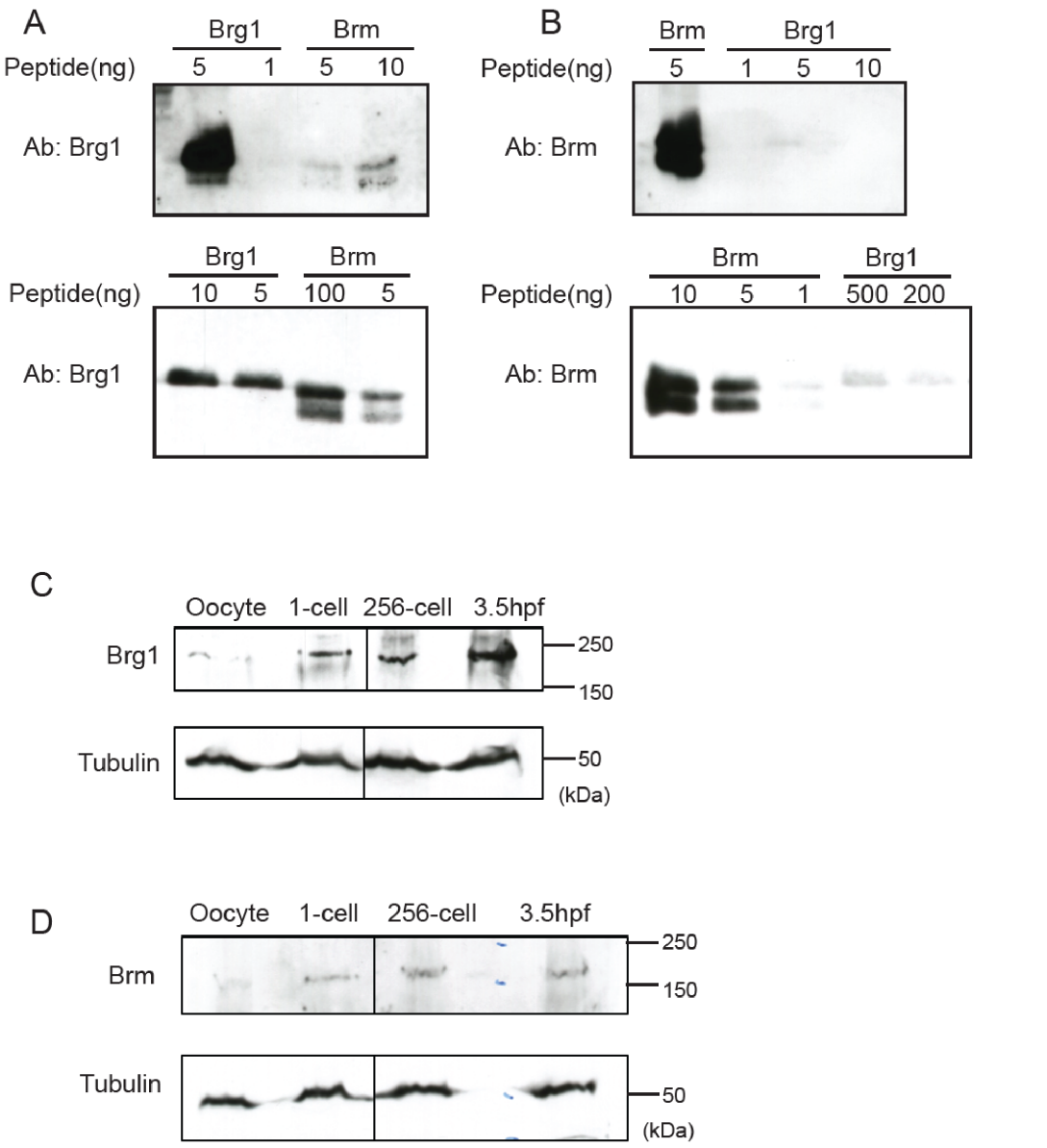


Figure 2.S2: Antibody specificity and protein level in zebrafish early embryo. (A) Immunoblotting using purified Brg1 antibody against antigen peptide of Brg1 and Brm. (B) Immunoblotting using purified Brm antibody against antigen peptide of Brm and Brg1. (C) (D) Immunoblotting of Brg1 and Brm in early zebrafish embryos, 30 embryos per each stage.

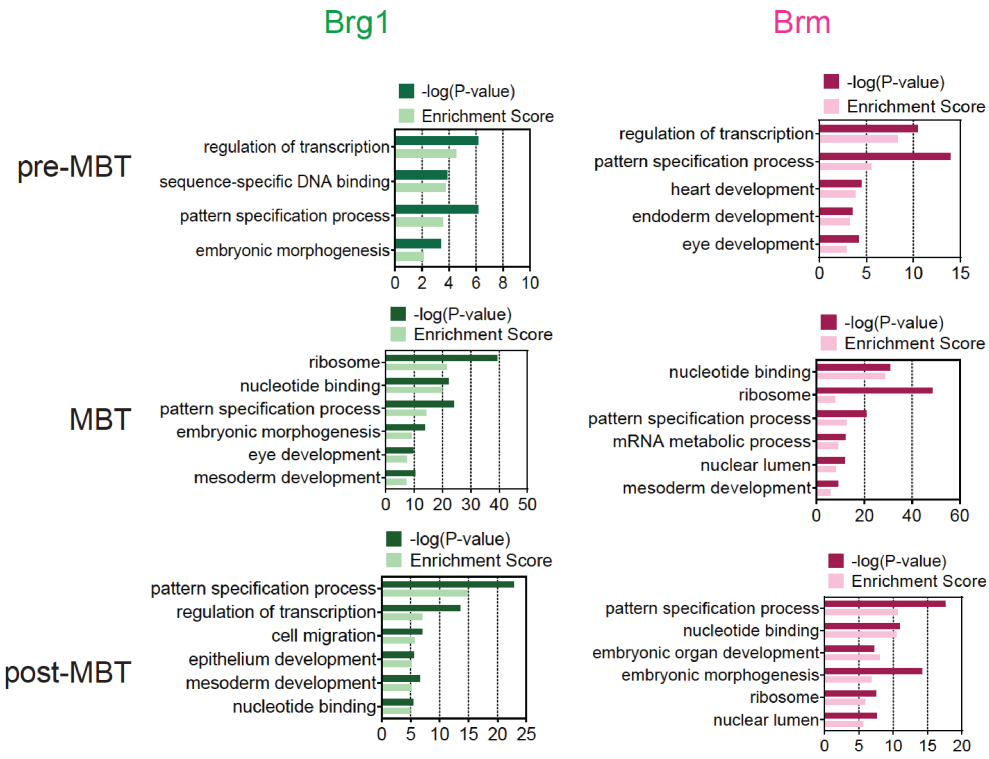
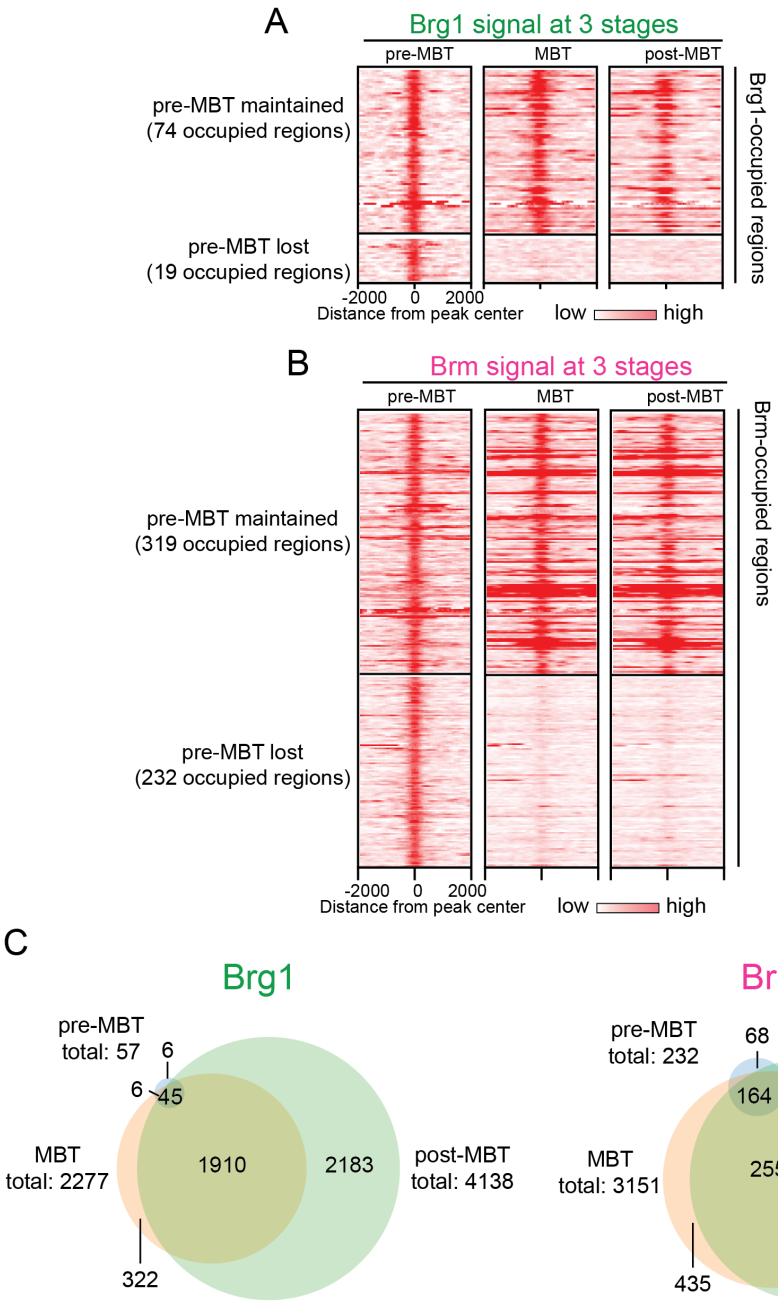


Figure 2.S3: Gene ontology analysis. Gene ontology analysis of Brg1- and Brm-occupied genes at each stage, analyzed by DAVID.

Figure 2.S4: Dynamic occupancy of Brg1 and Brm at each stage. (A) Zoom-in view of Brg1-occupied regions at pre-MBT. Heatmaps of Brg1 signal at each stage around peak centers ($\pm$2kb). (B) Zoom-in view of Brm-occupied regions at pre-MBT. Heatmaps of Brn signal at each stage around peak centers ($\pm$2kb). (C) Venn diagram analysis of the overlap between Brg1-occupied genes at three stages (left) and the overlap between Brm-occupied genes at three stages (right).


A
Brg1 signal at 3 stages
pre-MBT
MBT
post-MBT
pre-MBT maintained
(74 occupied regions)
pre-MBT lost
(19 occupied regions)
Brg1-occupied regions
-2000
0
2000
Distance from peak center
low
high
B
Brm signal at 3 stages
pre-MBT
MBT
post-MBT
pre-MBT maintained
(319 occupied regions)
pre-MBT lost
(232 occupied regions)
Brm-occupied regions
-2000
0
2000
Distance from peak center
low
high
C
Brg1
pre-MBT
total: 57
6
6
45
MBT
total: 2277
1910
2183
post-MBT
total: 4138
322
Brm
pre-MBT
total: 232
68
164
MBT
total: 3151
2552
1263
post-MBT
total: 3979
435

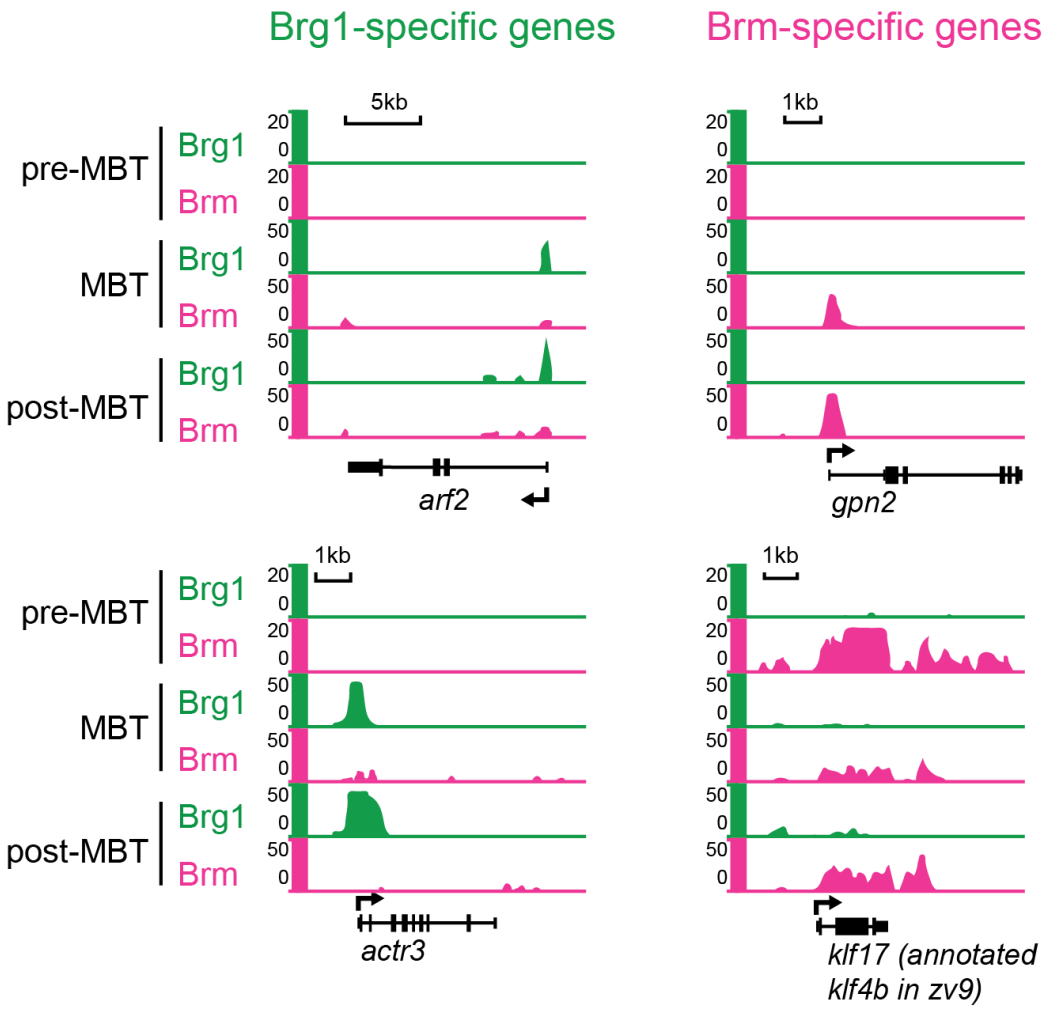
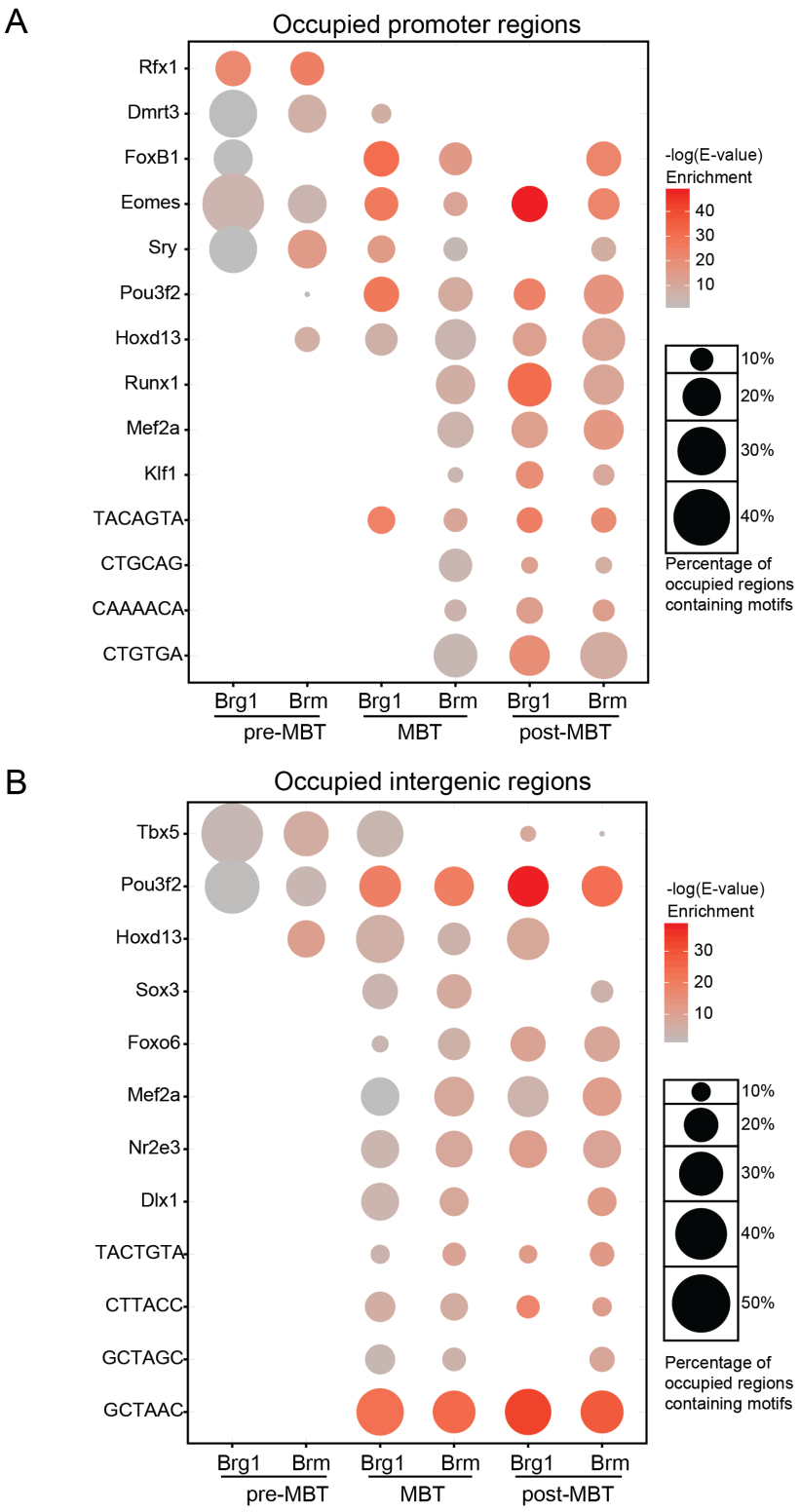


Figure 2.S5: Snapshots of dynamic occupancy of Brg1- and Brm-specific regions in early development.

Figure 2.S6: Motif analysis of Brg1- and Brm-occupied regions at intergenic regions and promoters. Regions that reside at intergenic regions (A), and promoters (B). The color of the circles represents the enrichment P-value. The size of the circle represents the percentage of occupied regions containing specific motif.


A
Occupied promoter regions
Rfx1
Dmrt3
FoxB1
Eomes
Sry
Pou3f2
Hoxd13
Runx1
Mef2a
Klf1
TACAGTA
CTGCAG
CAAAACA
CTGTGA
Brg1
Brm
pre-MBT
MBT
post-MBT
-log(E-value)
Enrichment
40
30
20
10
10%
20%
30%
40%
Percentage of
occupied regions
containing motifs
B
Occupied intergenic regions
Tbx5
Pou3f2
Hoxd13
Sox3
Foxo6
Mef2a
Nr2e3
Dlx1
TACTGTA
CTTACC
GCTAGC
GCTAAC
50%

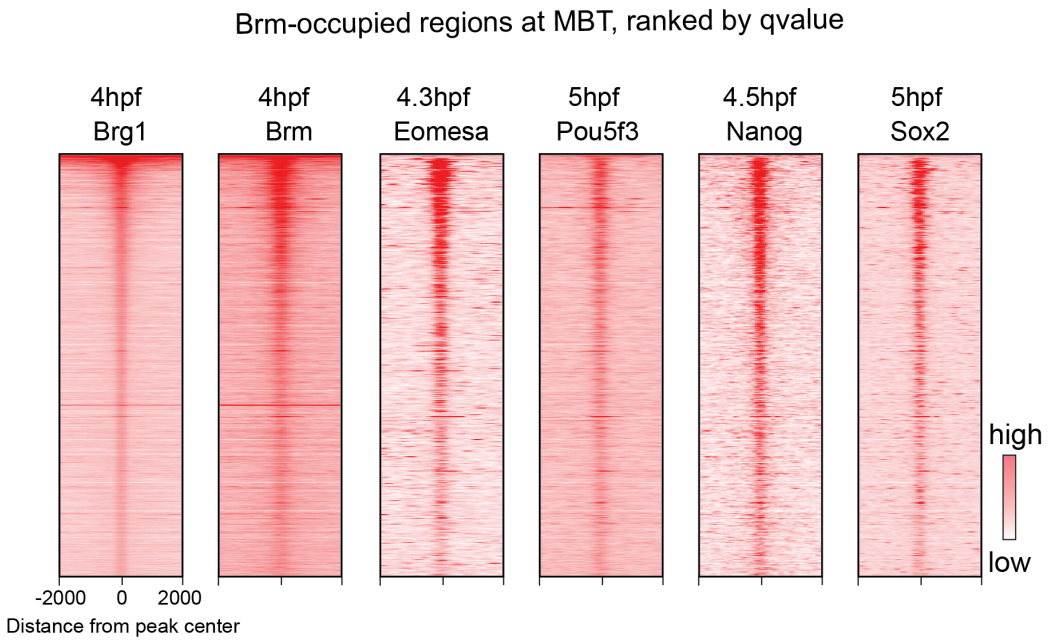
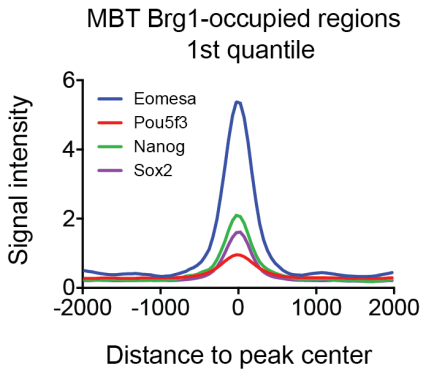
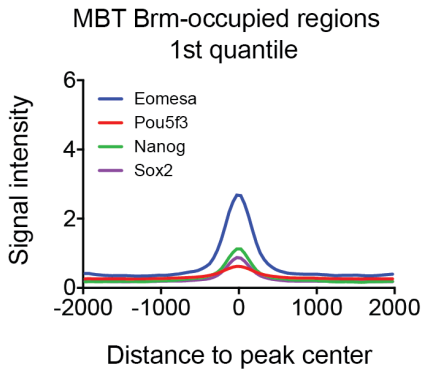
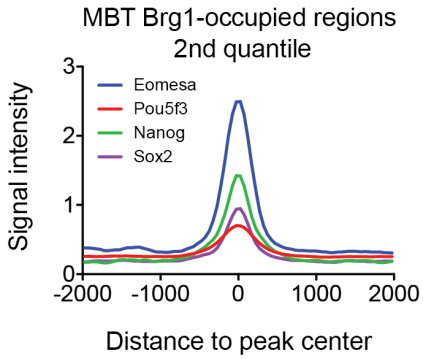


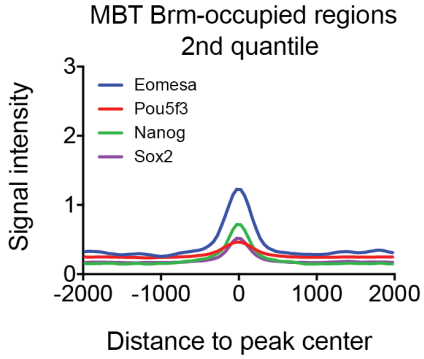
Figure 2.S7: Heatmaps of correlation between transcription factors and Brm. ChIP-seq signal (from left to right Brg1, Brm, Eomesa, Sox2, Pou5f3 and Nanog) around Brm-occupied regions at MBT, ranked by q-value.

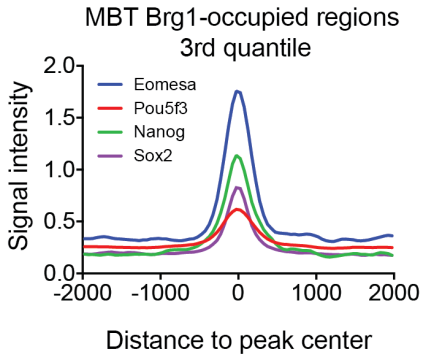
Figure 2.S8: Stronger Brg1- and Brm-occupied regions have higher enrichment of TF binding signal. Brg1- and Brm-occupied regions at MBT are divided into four groups based on the q-value at the summit of each peak. Then mean profiles of Eomesa, Sox2, Pou5f3 and Nanog signal (ChIP-seq data) around different groups of peaks are plotted.

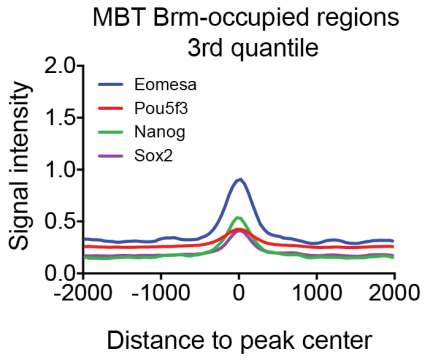

MBT Brg1-occupied regions
1st quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center

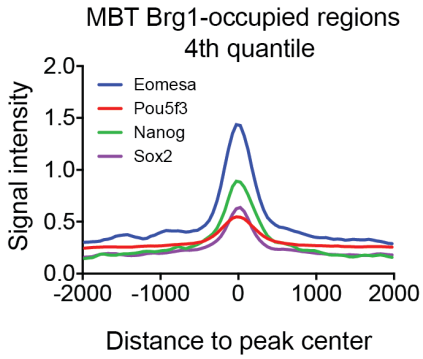

MBT Brm-occupied regions
1st quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center


MBT Brg1-occupied regions
2nd quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center


MBT Brm-occupied regions
2nd quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center


MBT Brg1-occupied regions
3rd quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center


MBT Brm-occupied regions
3rd quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center


MBT Brg1-occupied regions
4th quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center

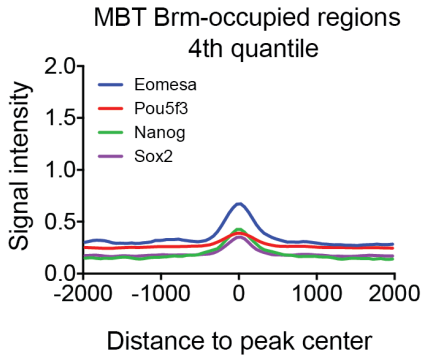
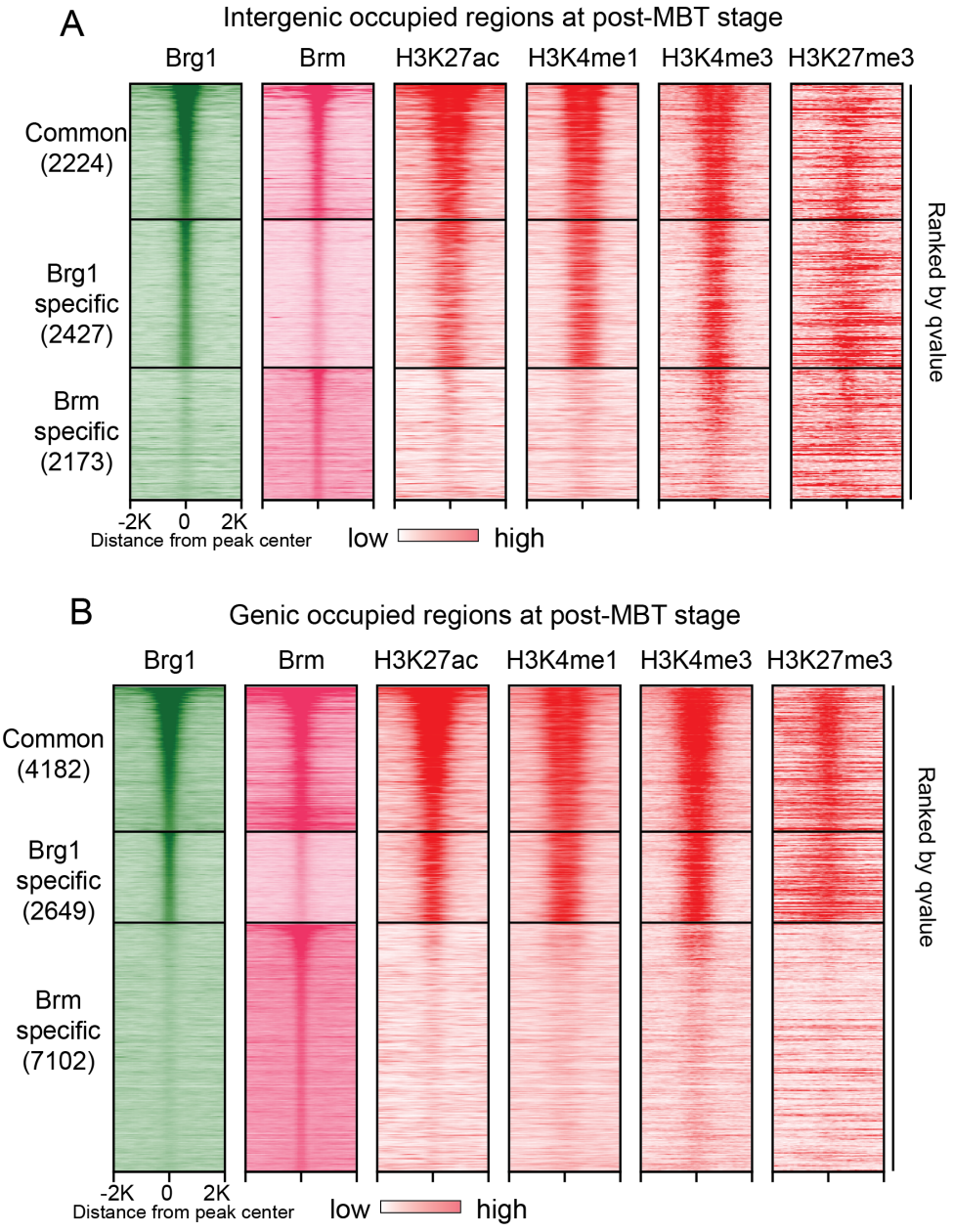

MBT Brm-occupied regions
4th quantile
Signal intensity
Eomesa
Pou5f3
Nanog
Sox2
Distance to peak center

Figure 2.S9: Snapshots of co-localization between transcription factors and Brg1/Brm. co-localization between Brg1, Brm and pluripotency at enhancers and promoters of pluripotency factors, where only eomesa is not bound by Brg1 and Brm.

Figure 2.S10: Brg1-occupied regions at post-MBT are enriched for active histone marks. (A) Heatmaps of Brg1, Brm, H3K27ac, H3K4me1, H3K4me3 and H3K27me3 signal (ChIP-seq) around occupied intergenic regions at post-MBT. Common: regions co-occupied by Brg1 and Brm; Brg1-specific: regions specifically occupied by Brg1; Brm-specific: regions specifically occupied by Brm. Each row represents $\pm$2kb around the center of occupied regions. Among each category, occupied regions are ranked by q-value. (B) Heatmaps of signals around occupied genic regions at post-MBT, as in (A).


A
Intergenic occupied regions at post-MBT stage
Brg1
Brm
H3K27ac
H3K4me1
H3K4me3
H3K27me3
Common (2224)
Brg1 specific (2427)
Brm specific (2173)
Ranked by qvalue
-2K 0 2K
Distance from peak center
low
high
B
Genic occupied regions at post-MBT stage
Brg1
Brm
H3K27ac
H3K4me1
H3K4me3
H3K27me3
Common (4182)
Brg1 specific (2649)
Brm specific (7102)
Ranked by qvalue
-2K 0 2K
Distance from peak center
low
high

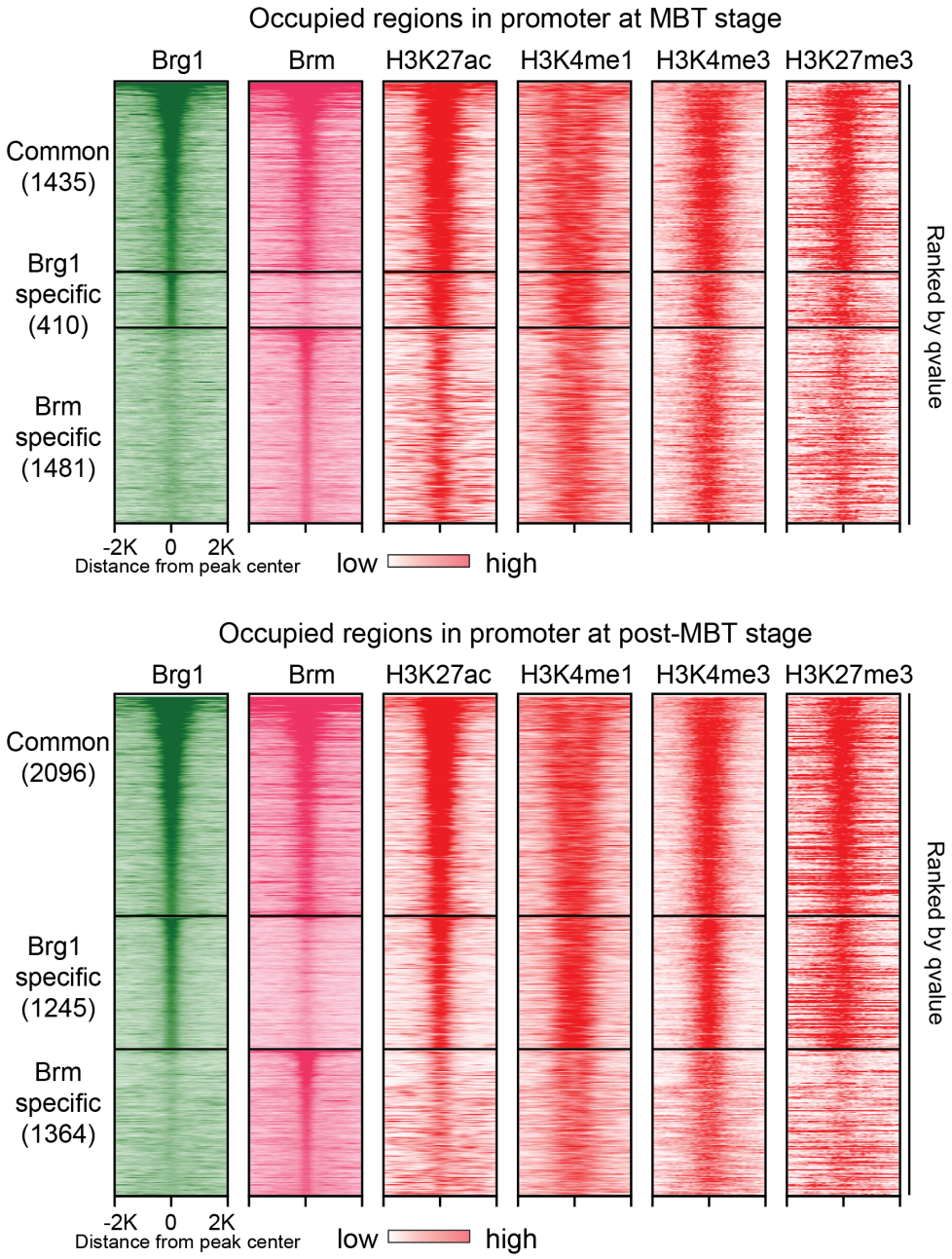


Figure 2.S11: Brg1-occupied regions at promoters are enriched for active histone marks. (A) Heatmaps of Brg1, Brm, H3K27ac, H3K4me1, H3K4me3 and H3K27me3 signal (ChIP-seq) around occupied promoter regions at MBT. (B) Heatmaps of signals around occupied promoter regions at post-MBT, as in (A).

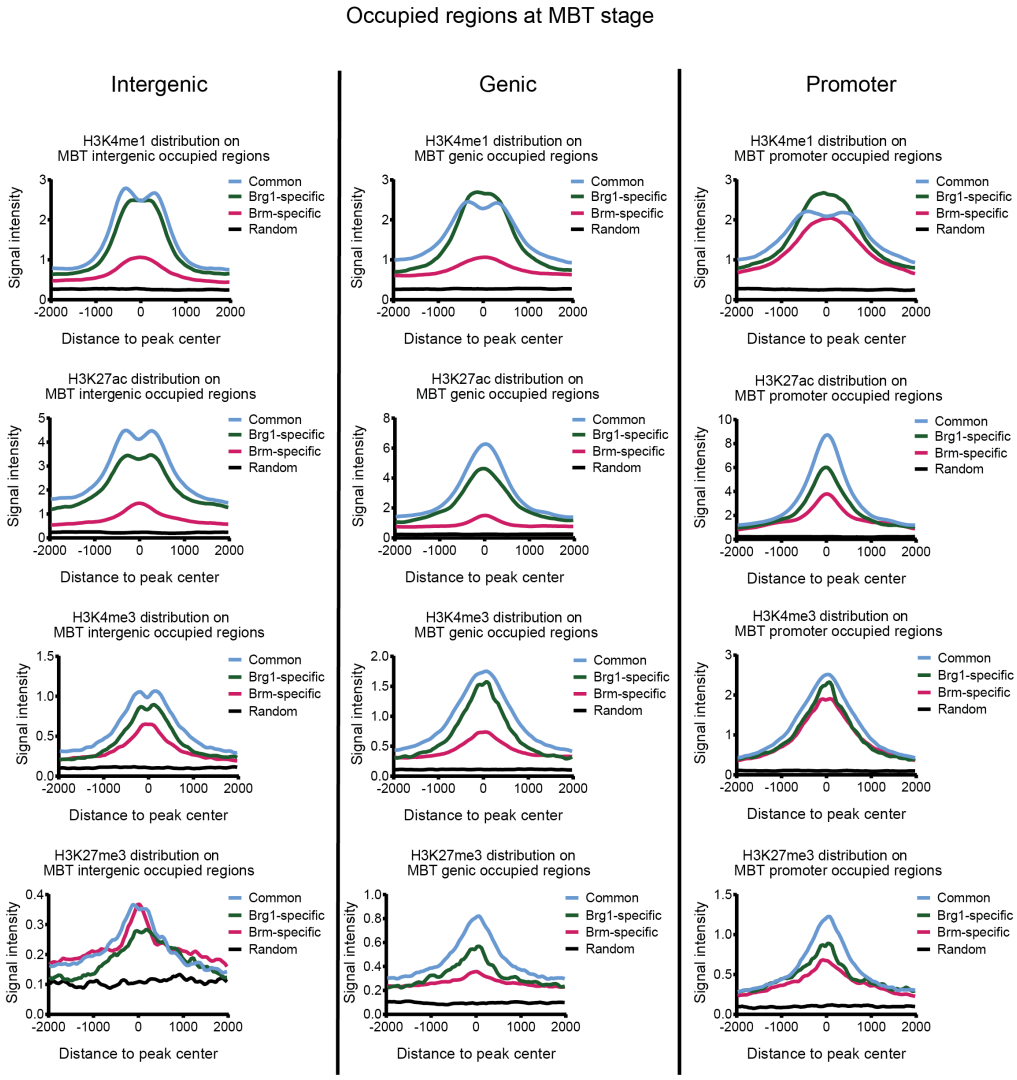


Figure 2.S12: Class average of histone modifications on commonly and specifically occupied regions at MBT stage. Occupied regions are classified into intergenic, genic and promoter regions, which are further grouped into common, Brg1-specific and Brm-specific regions. Next, the mean profiles of H3K27ac, H3K4me1, H3K4me3, H3K27me3 signal (ChIP-seq) were plotted.

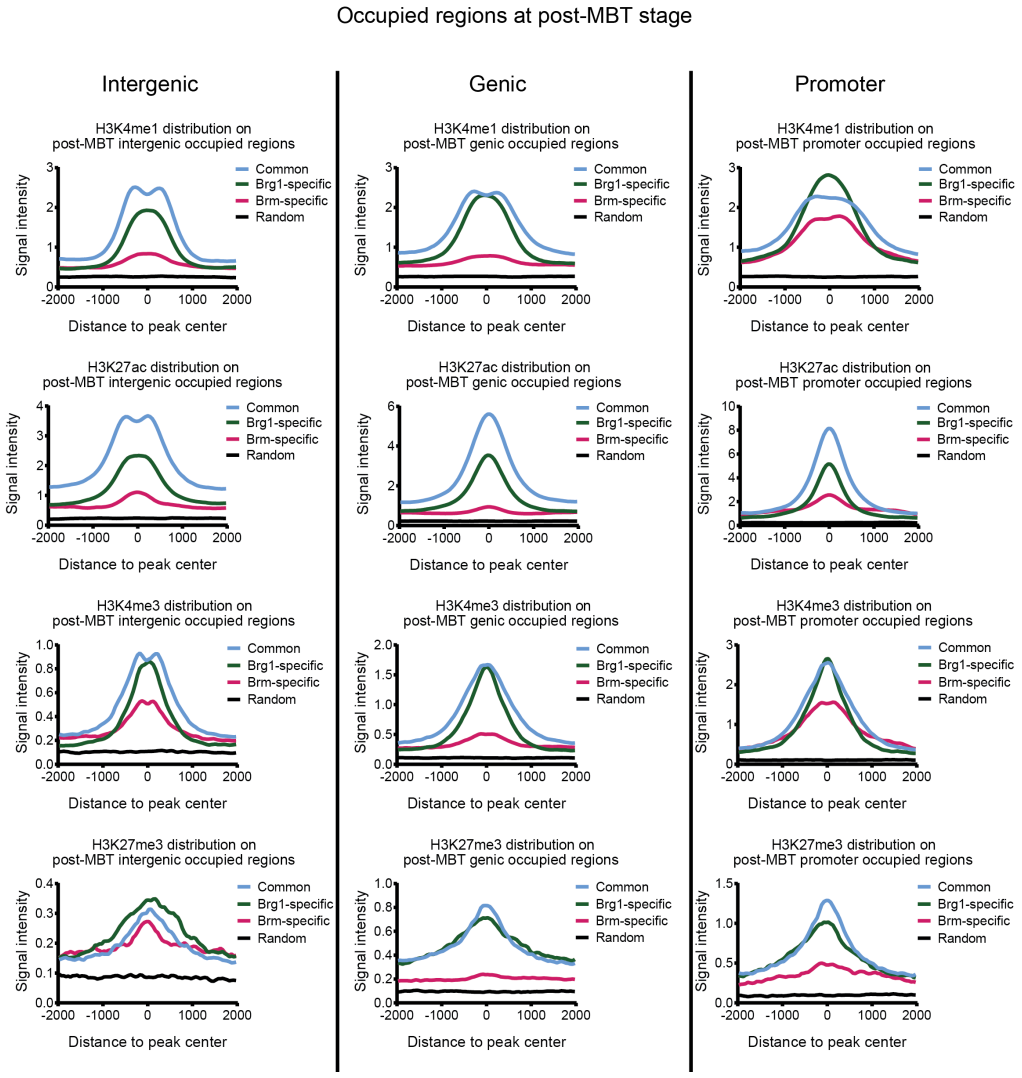


Figure 2.S13: Class average of histone modifications on commonly and specifically occupied regions at post-MBT stage. Occupied regions are classified into intergenic, genic and promoter regions, which are further grouped into common, Brg1-specific and Brm-specific regions. Next, the mean profiles of H3K27ac, H3K4me1, H3K4me3, H3K27me3 signal (ChIP-seq) were plotted.

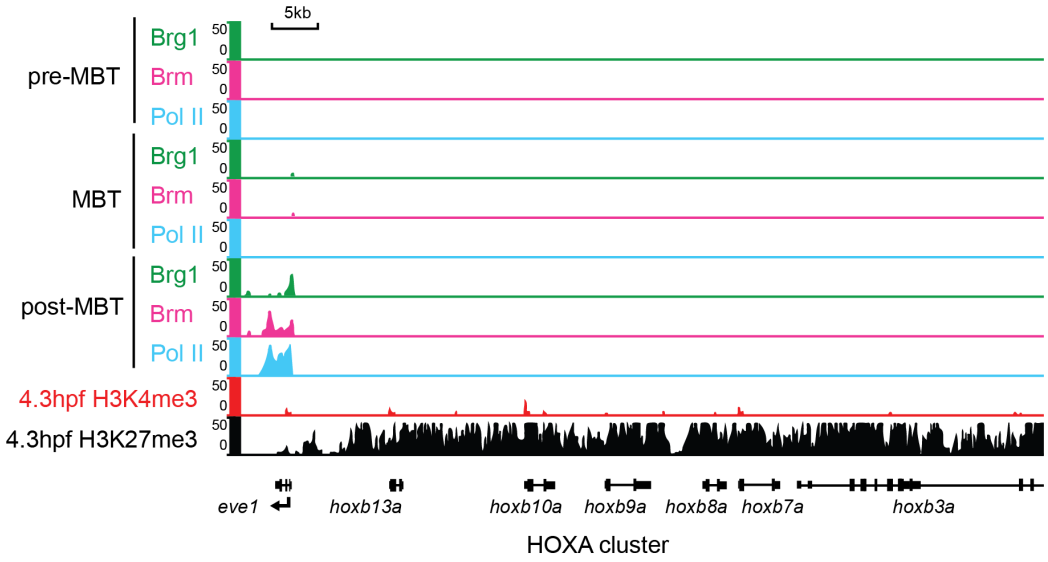


Figure 2.S14: Genome browser view showing Brg1, Brm and Pol II are absent at HOXA locus.

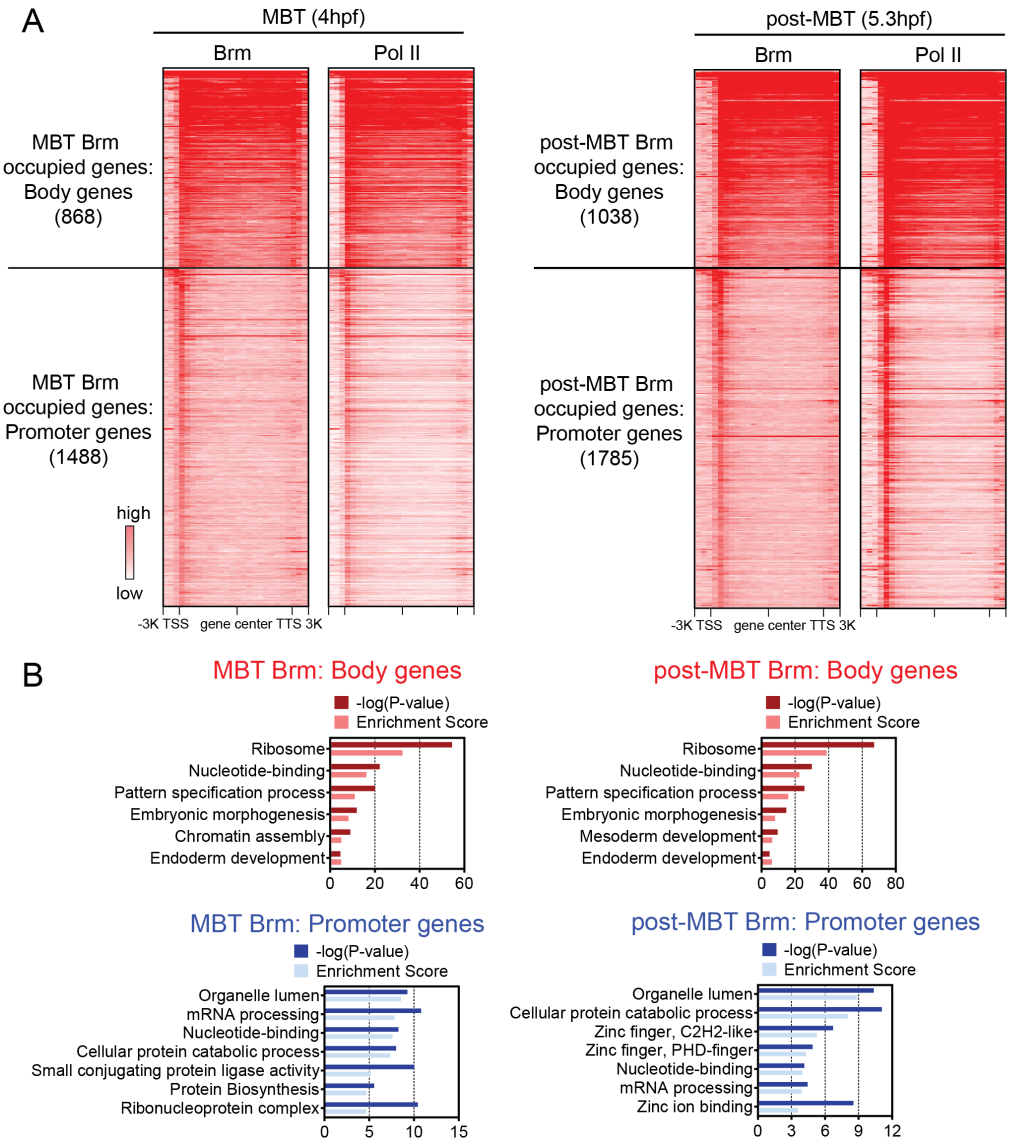
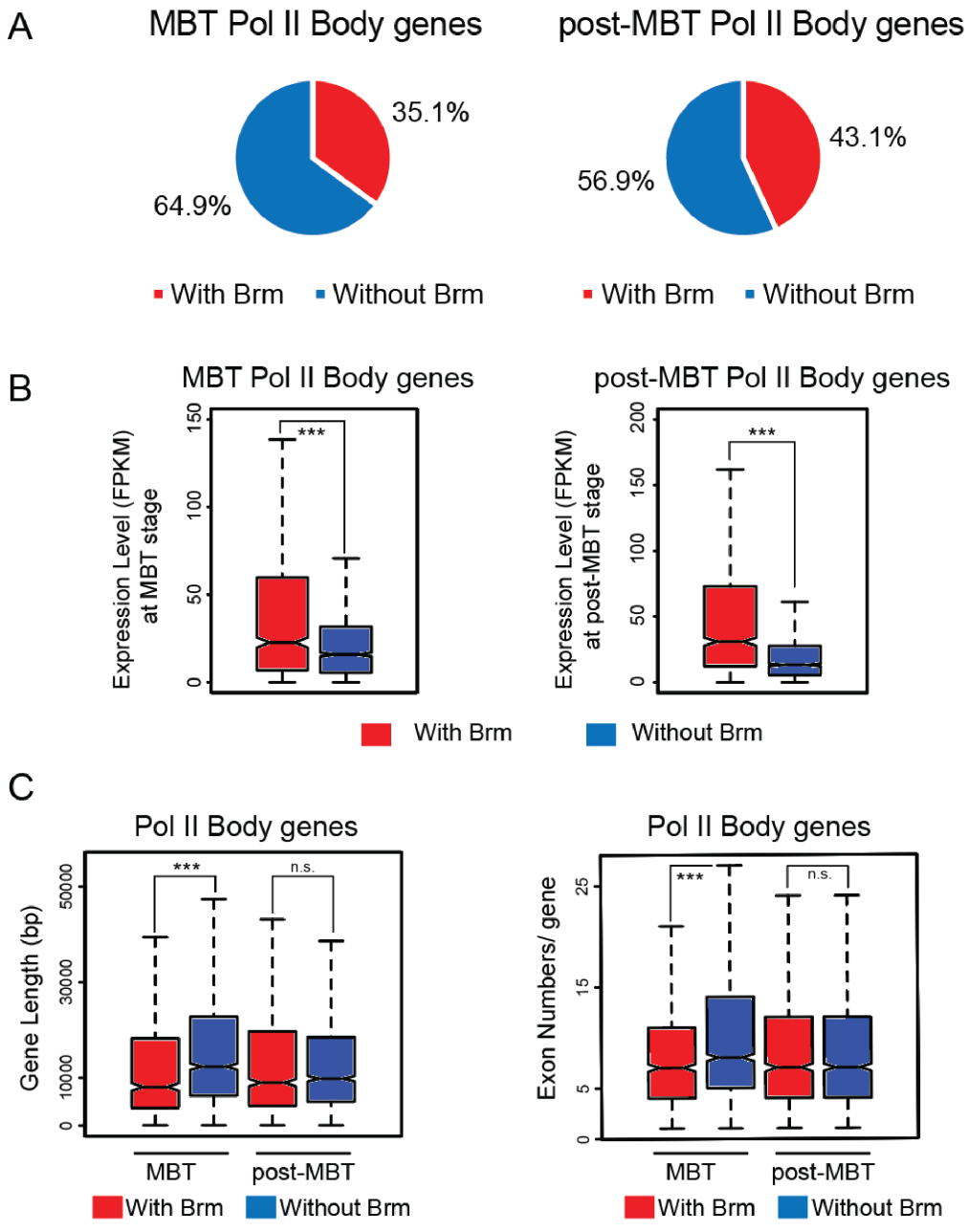


Figure 2.S15: Correlation of Brm and Pol II. (A) Heatmaps of Brm and Pol II ChIP-seq signal on Brm Body genes and Promoter genes at MBT (left) and, on those at post-MBT (right). Each row represents one gene, which includes 20 bins covering gene start to stop, 3 bins for upstream 3kb and 3 bins for downstream 3kb. (B) Gene Ontology of Brm Body genes and Promoter genes at MBT and post-MBT, analyzed by DAVID.

Figure 2.S16: A subset of Pol II Body genes is bound by Brm. (A) Pie charts of percentage of Pol II body genes bound by Brm at MBT and post-MBT. (B) Boxplot of expression level of MBT and post-MBT Pol II Body genes bound or unbound by Brm. (D) Boxplot of gene length and exon numbers of MBT and post-MBT Pol II Body genes bound or unbound by Brm. The difference between two groups is statistically significant by t test (*** denotes $p < 0.001$).


A
MBT Pol II Body genes
35.1%
64.9%
With Brm
Without Brm
post-MBT Pol II Body genes
43.1%
56.9%
With Brm
Without Brm
B
MBT Pol II Body genes
Expression Level (FPKM) at MBT stage
0
50
100
150
***
post-MBT Pol II Body genes
Expression Level (FPKM) at post-MBT stage
0
50
100
150
200
***
With Brm
Without Brm
C
Pol II Body genes
Gene Length (bp)
0
10000
30000
50000
***
n.s.
MBT
post-MBT
With Brm
Without Brm
Pol II Body genes
Exon Numbers/ gene
0
5
15
25
***
n.s.
MBT
post-MBT
With Brm
Without Brm

CHAPTER 3

THE LANDSCAPE OF ACCESSIBLE CHROMATIN IN EARLY ZEBRAFISH EMBRYO

3.1 Abstract

The early embryo undergoes extensive chromatin reprogramming to convert two terminally-differentiated gametes into a totipotent zygote. Chromatin accessibility is crucial for gene regulation and early embryo plasticity. However, the dynamics of accessibility and how it regulates early embryogenesis remain unknown. Here, we provide the global landscapes of accessible chromatin at three embryonic stages and adult liver determined by ATAC-seq. Chromatin becomes gradually accessible during early embryogenesis, with a substantial increase from pre-MBT to MBT. At MBT, more than half of open chromatin regions are distributed around genes and are highly enriched at promoters. In general, transcription of genes correlates with open chromatin, except a small percentage of genes. Interestingly, the promoters of these genes are enriched for H3K4me1 and H3K27me3, which is a new bivalent mark identified at promoters. In addition, intergenic open chromatin regions displayed various patterns of H3K4me1 and H3K27ac. I further examined Pol III genes, which behave opposite to Pol II: Pol III genes are generally open pre-MBT, but become inaccessible as the embryo develops. Furthermore, both active histone marks and Pol II are enriched around the TSS of most 'open' Pol

III genes at MBT. Importantly, candidate transcription factors were identified at open chromatin regions based on motif analysis, including pluripotency factors Pou5f3, and Sox2, and novel factors Nr2e3, NFYB and others. Finally, nucleosome positioning can be profiled utilizing ATAC-seq fragments with length of 120-220bp.

3.2 Introduction

Upon fertilization, the embryo undergoes extensive chromatin reorgranization in order to acquire the ability to differentiate into any cell lineage. Specific transcription circuits are the basis of pluripotency and subsequent cell fate specification, which is orchestrated at multiple levels – including DNA methylation, histone modifications, chromatin remodeling, as well as higher-order chromatin structure (Burton & Torres-Padilla, 2014). How chromatin profiles are established and how chromatin profiles instruct gene expression during embryogenesis are pivotal questions that remain to be answered.

A key aspect of the chromatin profile is chromatin accessibility. DNA is wrapped around nucleosome which precludes binding from most nonhistone proteins, known as inaccessible chromatin. For gene activation, chromatin at promoters and/or enhancers has to become permissive. The accessible chromatin allows binding of transcription factors (TFs) and other co-activators, leading to the following engagement of polymerase and transcription (Levine & Tjian, 2003). Most TFs cannot bind nucleosomal DNA, except pioneer transcription factors, such as FOXA and GATA factors (Zaret & Carroll, 2011). These pioneer factors bind nucleosomal DNA and recruit chromatin remodelers to open up the chromatin and make it accessible to other transcription factors and co-activators. Profiling the global

chromatin accessibility allows us to identify the candidate pioneer factors, as well as key transcription factors that bind the opened chromatin and actively regulate gene expression.

Extensive work have focused on Pol II genes to understand the interplay between chromatin accessibility and gene expression. However, evidence suggests expression of Pol III genes is also modulated at the chromatin level (Hartley & Madhani, 2009). RSC complex is a SWI/SNF-family chromatin remodeler in yeast. Loss of the RSC function by mutating the catalytic subunit, Sth1, results in globally reduced transcription of Pol I, II and III genes. Interestingly, Pol III genes consist of a large class of RSC-occupied genes, and they have the lowest histone-occupancy genome-wide (Parnell et al., 2008). Discovered in HeLa and other cell lines, many active Pol III genes are in close proximity to Pol II genes who have permissive chromatin (Oler et al., 2010). Furthermore, Pol III occupancy is significantly correlated with active histone modifications including H2AZ and H3K4me3. Collectively, these data suggest Pol III gene activation also requires accessible chromatin and may take advantage of the active chromatin environment established by the chromatin remodeling occurring at Pol II genes (Oler et al., 2010).

In Chapter 2, we identified the regions bound by SWI/SNF complex, using Brg1 and Brm as a proxy. However, there are almost certainly other regions that either rely on other chromatin remodelers and/or are accessible due to other mechanisms, and will be missing if solely dependent on mapping the occupancy of Brg1 and Brm. Therefore, to gain a comprehensive understanding of chromatin landscape in early zebrafish development, we conducted ATAC-seq at three embryonic stages, and also in differentiated liver (for comparisons). Fragments were grouped

into ATAC-open and ATAC-mono based on length, which others have established represent open chromatin (lacking nucleosomes) and mono-nucleosomes, respectively. Notably, I found that focal chromatin accessibility increases as the embryo develops. These open chromatin regions are highly enriched at promoters, and correlated with gene expression level. Comparing open chromatin with histone modifications also revealed many interesting features at promoters and intergenic regions.

3.3 Results

3.3.1 ATAC-seq in early zebrafish embryos (and liver) reveals accessible and nucleosome-occupied regions

To determine the accessible chromatin regions and how they regulate early embryogenesis, we performed ATAC-seq experiments in three embryonic stages and one differentiated tissue, adult liver. Two replicates for each stage are submitted for paired-end sequencing, as well as Tn5 transposed naked DNA, which controls for sequence bias of Tn5 enzyme (Figure 3.1). We obtained around 100M reads for each replicate, which gave 10-fold genome coverage on average (Table 3.1). The two replicates for each stage were highly correlated, with Pearson Correlation score higher than 0.99 (Figure 3.2A-D). Notably, the samples from embryonic stages are clustered together and the liver sample is divergent from the embryonic samples (Figure 3.2E).

By plotting the fragment size distribution, I found there were three major groups of reads: the first group of reads expands from 0-100bp; reads from the second group mainly had length from 100 to 200bp; the third group had a much shallower peak and expanded from 300bp to 400bp (Figure 3.3A). Group 2 and

Group 3 reads displayed as bell curves, and their distributions were fitted by Gaussian Distribution. The derived mean and standard deviation were used to group reads with length from 120-220bp as ATAC-mono (group 2), and reads with length from 270-410bp as ATAC-di. We further classified reads with length (0-100bp) as ATAC-open. The three groups of reads likely represent mono-nucleosome, di-nucleosome and open chromatin, respectively. Particularly, ATAC-mono reads displayed periodical subpeaks corresponding to the helical pitch of DNA (Figure 3.3A, arrowhead). The percentage of ATAC-open reads in pre-MBT samples is the highest among all stages, and it gradually decreased as the embryo developed. In contrast, the percentage of ATAC-mono reads increased along development (Figure 3.4A). One example is shown in Figure 3.4B, where the ATAC-open and ATAC-mono clearly reflected two distinct populations, which were unresolved in the ATAC-total.

3.3.2 Substantial increase of chromatin accessibility from pre-MBT to MBT

Using MACS2 (FDR $<$ 0.001) with subsequent filtration using stringent cutoff (length $\geq$ 100bp and $-$log(q-value) $\geq$10), I identified 850, 24331, 32934 and 26799 strong open regions at pre-MBT, MBT, post-MBT and liver (Figure 3.5A). The numbers of open chromatin regions increased dramatically from pre-MBT to post-MBT, suggesting an increase in chromatin accessibility in early development. Interestingly, open chromatin regions decreased from post-MBT to liver, indicating that differentiated tissue has fewer open chromatins compared to embryonic cells. Of note, very few open chromatin regions were identified at pre-MBT; this feature could be caused by technical difficulties. The chromatin is largely condensed

during mitosis in the fast cell cycle (15 min) at pre-MBT, and thus Tn5 enzyme could inefficiently transposase in the open chromatin regions. Alternatively, this might reflect a true chromatin landscape at pre-MBT that is the chromatin is refractory globally, which is consistent with two recent studies investigating the open chromatin in mouse embryo (Lu et al., 2016; Wu et al., 2016). More investigations are needed to determine which is the case.

To characterize the open chromatin regions, I examined their distribution on different genome features. Although a majority of pre-MBT open regions localized at the intergenic area, open regions at MBT, post-MBT and liver were highly enriched at promoters and coding exons (Figure 3.5B). Next, I analyzed which regions at one stage were lost or maintained in the next stage (Figure 3.5C). For instance, $\sim$70% of pre-MBT peaks were lost at MBT, while $\sim$30% were maintained at MBT. Examining the genome distribution of the lost peaks revealed that open regions at intergenic and coding area were preferentially lost compared to the regions at promoters (compare Figure 3.5D to Figure 3.5B). In addition, substantial regions were gained at MBT, post-MBT and liver, compared to their previous stages (Figure 3.5E). MBT-gained regions which were greatly enriched at promoter and coding regions, whereas, liver-gained regions have a higher fraction in intron and intergenic area (Figure 3.5F). This may suggest more enhancers become accessible in liver compared to MBT.

To understand what the functions of genes associated with open chromatin are, I performed Gene Ontology analysis. For gained regions at MBT, ‘transcription’ and ‘nucleotide binding’ were the most enriched terms. At post-MBT, enrichment additionally included ‘pattern specification process’ and development terms like

'eye development', and 'determination of symmetry'. As expected, gained open regions in liver were significantly enriched at various enzymatic processes that are specific for liver (Figure 3.6), while transition from post-MBT embryo to liver was accompanied by the loss of transcription and development related terms (Figure 3.6). The distinct GO terms at each stage reflect that open regions were associated with different genes at each stage, and a few examples were provided in Figure 3.7. *sp3a* and *actb2* are genes ubiquitously transcribed and have open chromatin across MBT, post-MBT and liver, but not in control (DNA), whereas *apoa* and *serpina1* are key regulators in liver, and exhibited open chromatin specifically in liver. Additionally, *foxh1* and *pou5f3* are genes required for early development and were only open at MBT and post-MBT stages. In summary, our ATAC-seq data reveal a highly dynamic chromatin in early development, and also demonstrate the embryonic chromatin profiles are distinct from that of differentiated tissue.

3.3.3 Genes associated with open chromatin have higher expression

Around 60% of open chromatin regions at MBT, post-MBT and liver resided in the gene-related area (Figure 3.5B), which was associated with 23, 8356, 9846, and 8936 genes at pre-MBT, MBT, post-MBT and liver (Figure 3.8A). Next, Meta-gene analysis was performed to determine where open chromatin regions were enriched related to genes. Open-chromatin signals were highly enriched at promoters of occupied genes at MBT, post-MBT and liver, and enriched to a lesser extent at pre-MBT. Moreover, open-chromatin signals of MBT and post-MBT were moderately enriched at the downstream regions (Figure 3.8B). Open chromatin at promoters allows transcription factors and Polymerase to bind, and usually in-

dicates active transcription. Therefore, I asked whether genes associated with open chromatin have elevated transcription. Indeed, genes associated with open chromatin at MBT and post-MBT have higher expression compared to genes that do not have open chromatin (Figure 3.8C).

To interrogate whether gene expression correlates with open chromatin accessibility, I ranked all genes based on their expression at either MBT (Figure 3.9A) or post-MBT (Figure 3.9C). As expected, highly transcribed genes (on top) have stronger open chromatin signal. A similar conclusion can be drawn, when genes were classified into 4 quantiles based on expression (1st quantile indicated the top 25% transcribed genes) (Figure 3.9B, D). These data indicate highly transcribed genes are associated with strong open chromatin.

3.3.4 Potential new bivalent mark at promoters

Among all genes associated with open chromatin at MBT, the majority (78%, 6571 out of 8356 genes) of which have open chromatin at promoters, and are actively transcribed (FPKM $>$1). However, $\sim$16% (1041 out of 6571) of genes barely have any transcription (FPKM $<$1) even they are associated with open chromatin regions. I then sought features that might explain why this subset of genes have accessible promoters but are not transcribed. Firstly, I ranked all ATAC-open peaks that resided at promoter based on the expression level of their associated genes. Next, I profiled $-$log(q-value) of ATAC-open signals, ChIP-seq signals of H3K27ac, H3K4me1, H3K4me3 and H3K27me3 around these peaks (Figure 3.10A) (Zhang et al., 2014; Bogdanović et al., 2012). Interestingly, regardless of the expression level of associated genes, open chromatin regions

exhibited similar strength (indicated by the intensity of color in the green heatmap of Figure 3.10A). This ruled out the possibility that those genes have a weaker open chromatin, which is insufficient to activate transcription. Next, I focused on histone modifications. As expected, H3K27ac and H3K4me3 were highly correlated with gene expression, while H3K27me3 are enriched for genes that are not transcribed. Most interesting, genes down on the ranking list had high levels of H3K27me3 and H3K4me1, and void of H3K4me3 and H3K27ac.

To investigate this further, I selected the top 10% peaks (associated gene are highly transcribed) and the bottom 10% peaks (associated gene are not transcribed). Notably, the top 10% peaks have high level of active histone marks (H3K27ac, H3K4me1, and H3K4me3) but not repressive mark (H3K27me3) (Figure 3.10C). In striking contrast, the bottom 10% genes have a much lower H3K27ac signal, and instead displayed high occupancy of H3K27me3 and H3K4me1 (Figure 3.10C). The active H3K4me1 and repressive H3K27me3 form a new bivalent mark, which is different from the canonical bivalent mark defined by H3K4me3 and H3K27me3 (Bernstein et al., 2006). I speculate the new bivalent mark might have a similar function to the canonical one, which is to poise genes for expression at a later stage. Whether this new bivalent mark is specific to zebrafish, or specific to early embryo need to be further explored.

3.3.5 Intergenic open chromatin regions at MBT

A number of the ATAC-open peaks ($\sim$40%) were distributed at intergenic regions, which could be indicative of putative enhancers. Only ChIP-seq data of histone modifications (H3K27ac, H3K4me1, H3K4me3 and H3K27me3) at 4.3hpf

(close to 4hpf, MBT) are public available. Therefore, I focused on ATAC-open intergenic regions at MBT, and profiled the signal of histone modifications around the intergenic peaks. As expected, H3K27ac and H3K4me1, histone modifications that mark enhancers, were enriched at those intergenic regions. In contrast, no significant occupancy of H3K4me3 and H3K27me3 was observed at those loci (Figure 3.11A). To determine whether the intergenic regions have different combination of histone marks, I performed *k*-mean clustering with the normalized H3K27ac and H3K4me1 signal. Notably, a large percentage of the intergenic regions were co-bound by H3K27ac and H3K4me1, while some regions were preferentially bound by H3K27ac (Figure 3.11B, group 1 and group 2) or H3K4me1 (Figure 3.11B, group 3). Group 4 represented the regions that were only marked by H3K4me1, and presumably were inactive enhancers. Interestingly, around 40% of intergenic regions are nearly lacking H3K4me1 and H3K27ac (Figure 3.11B, group 5). Whether other histone modifications mark these intergenic regions will be interesting to know. In summary, the majority of the intergenic open chromatin regions are associated with at least one active histone modifications, and could be putative enhancers.

3.3.6 The number of Open Pol III genes decreases as the embryo develops

Expressions of Pol III genes are regulated at the chromatin level as well. To interrogate whether Pol III genes also associated with open chromatin in the early embryo, I first identified all predicted Pol III genes based on zv10 genome annotation (Table 2). tRNAs were identified from repeat masked tracks, and only those longer than 72bp were kept as potential tRNAs. For all the Pol III genes, ATAC-open

signal at each stage was collected around the TSS, and was subjected to *k*-mean clustering. Interestingly, a lot more Pol III genes are relatively open at pre-MBT compared to that at MBT, post-MBT and liver (Figure 3.12A). Across all stages, a subset of Pol III genes has a very strong ATAC-open signal, and were referred as 'Open Pol III genes'. Notably, 1354 Pol III are potentially open at pre-MBT, and this number decreases as the embryo develops (Figure 3.12B). Particularly, nearly 1000 Pol III genes lost the open chromatin signal from MBT to post-MBT. This suggests that Pol III genes are of greater need in early embryo (pre-MBT and MBT), potentially to facilitate the translation of the vast amount of transcripts newly generated at zygotic genome activation.

I then focused on the Open Pol III genes at MBT. The defined Open Pol III genes indeed have more ATAC-open signal at MBT, which was reduced at post-MBT (Figure 3.13A). Interestingly, these genes also have higher active histone modifications, particularly H3K27ac, whereas repressive mark, H3K27me3, was nearly absent (Figure 3.13B). Previous work from our lab suggested active Pol III genes are close to Pol II genes that usually have an open chromatin (Oler et al., 2010). In agree with this observation; around 16% of Open Pol III genes was in close proximity within Pol II genes (Figure 3.13C). Moreover, Pol II was highly enriched at the TSS of open Pol III genes, and moderately enriched at the regions around TSS. In contrast, Pol II was nearly absent in the 2kb regions around TSS of Pol III genes that are not open at MBT (Figure 3.13D). To sum, I have identified Pol III genes associated with open chromatin at each stage. The number of Open Pol III genes decreased dramatically from MBT to post-MBT. Additionally, promoters of Open Pol III genes were enriched for active histone marks and Pol II occupancy.

3.3.7 Motif analysis of open chromatin regions at different stages

Transcription factors may serve as 'pioneer factors' to facilitate the establishment of open chromatin, or can be targeted to the accessible chromatin for subsequent gene regulation. Therefore, investigating the over-represented motifs at open chromatin regions is informative for identification of candidate transcription factors involved in both aspects. To this end, I employed Regulatory Sequence Analysis Tools (RSAT) to determine the enriched motifs at open chromatin regions of four stages (Figure 3.14) (Medina-Rivera et al., 2015). At pre-MBT peaks, only three motifs were identified, which were all presented in nearly 80% of all pre-MBT peaks. However, the enrichment was barely significant (E-value $<$0.05), which was largely due to the small number of peaks identified at pre-MBT. Among these, Nr2e3 is of particular interest, as its homolog, Nr5a2, was highly enriched at open chromatin regions in mouse pre-implantation embryo (Figure 3.14A)(Wu et al., 2016). Knockout of Nr5e2 leads to embryo lethality at E6.5, and reduced the expression of pou5f1 and Nanog (Wu et al., 2016). Altogether, they suggest Nr2e3 might play an important role in zebrafish early development that is previously unknown. Another interesting motif is NFYB, which was highly enriched at MBT, post-MBT and liver (Figure 3.14A). This motif was also enriched in early mouse embryos, and knockdown of Nfya arrested mouse embryo at morula stage (Lu et al., 2016). These data suggest the regulation of early embryogenesis between zebrafish and mouse might be similar. Furthermore, POU5F1 and SOX motifs were specifically enriched at MBT and post-MBT embryos, but not in liver (Figure 3.14A). This not only confirmed the identified open chromatin regions were cell-type specific, but also support the known importance of zebrafish Pou5f1 (known

as Pou5f3 in zv10) and SOX factors.

To determine whether the candidate transcription factors at enhancers and promoter are different, I partitioned open chromatin peaks into promoter peaks and intergenic peaks, and performed the same motif analysis (Figure 3.14B, C). Additional motifs were discovered in intergenic area (i.e. RARG, GATA4) and promoter area (i.e. MNT, HOXA13). Intriguingly, many motifs were co-enriched at the promoter peaks of MBT, post-MBT and liver, which was in contrast to the variability observed at intergenic area. This is in line with the notion that enhancers are more dynamic compared to promoters. Of note, CTCF motif was only enriched at the intergenic area but not at promoter, which is consistent with its role as insulator. Finally, I also identified liver specific motifs including USF2, ELK4, NFIC, and GATA4. Taken together, I reveal diverse enriched motifs at different stages. The corresponding factors to some motifs (i.e. POU5F1, SOX2) are known to regulate zebrafish early development, whereas, factors of other motifs (i.e. NFYB, Nr3e2, YY, CREB1) have not been studied for their roles in zebrafish early development, and will be candidates for future work.

3.3.8 Eomesa, Sox2, Pou5f3 and Nanog are enriched at open chromatin regions

Among the enriched motifs at MBT and post-MBT, I am particular interested in Eomes, Pou5F1, and Sox motifs; Eomes promotes trophectoderm differentiation in mammals, while Pou5f1 and Sox factors promote pluripotency and ES-cell self-renewal, and early zebrafish development (Russ et al., 2000; Niwa, 2007; Leichsenring et al., 2013). To interrogate whether the corresponding proteins indeed play roles in the regulation of chromatin accessibility, I profiled the public ChIP-seq

data of Eomesa (4hpf), Sox2 (5hpf), Pou5f3 (5hpf) and Nanog (4.3hpf) around the open chromatin regions (Figure 3.15A-F) (Nelson et al., 2014; Leichsenring et al., 2013; Xu et al., 2012). Notably, all four factors were enriched at the centers of open chromatin regions at MBT and post-MBT. In addition, Eomesa have higher enrichment at intergenic peaks compared to promoter peaks, whereas Nanog, Sox2 and Pou5f3 have similar intensity at intergenic and promoter peaks.

To understand whether Eomesa, Sox2, Nanog and Pou53 function individually or cooperatively, ChIP-seq signals of Eomesa, Sox2, Nanog and Pou53 were normalized, combined and subjected to *k*-mean clustering. I observed some intergenic regions were bound by all four TFs (Figure 3.15G, group 1), while others were preferentially bound by Nanog (Figure 3.15G, group 4). Interestingly, group 2 and group 3 peaks showed reciprocal binding pattern, where group 2 peaks were bound by Eomesa and Nanog, and group 3 peaks were bound by Sox2 and Pou5f3 (Figure 3.15G, group 2 and 3). In addition, most of intergenic peaks ($\sim$70%) were not bound by the four TFs.

In contrast to the intergenic peaks, most promoter peaks were bound by individual or a combination of TFs (Figure 3.15H). A small percentage of promoter peaks were bound by all four factors (Figure 3.15H, group 1), while other peaks were predominantly associated with Pou5f3 (Figure 3.15H, group 5). Group 4 peaks seem associated with Sox2, Pou5f3 and Nanog. Intriguingly, peaks from group 2 or group 4 are preferentially bound by Eomesa or Nanog, respectively. Collectively, these data unveil Eomesa, Sox2, Nanog and Pou5f3 are enriched at open chromatin regions, and form different regulatory networks at promoters and intergenic area.

3.3.9 Pou5f3 may serve as pioneer factor for the establishment of open chromatin

Transcription factors in general are unable to bind DNA wrapped with histones, except a special class of 'pioneer factor' including FoxA and GATA factors (Zaret & Carroll, 2011). By recruiting chromatin remodeler, such pioneer factors can open the local chromatin, which becomes accessible to other transcription factors. Since Pou5f3, Sox2, Nanog are key regulators of zygotic genome activation in zebrafish, it is intriguing to known whether they could also serve as pioneer factors. To test, I collected ATAC-open signals at MBT around the ChIP-seq binding sites of Sox2, Pou5f3 and Nanog, as well as Eomesa, followed by *k*-mean clustering. For all factors, only 20-30% of peaks have strong open chromatin signals (Figure 3.16A, B), which indicates that these four factors have the ability to bind closed chromatin. In particular, nearly 80% of Pou5f3 binding sites barely exhibited low ATAC-open signal, which argues strongly that Pou5f3 could be a pioneer factor.

If pioneer factors are able to open up the local protein, the nucleosome density should be decreased. To investigate whether these factors (Eomesa, Pou5f3, Sox2, Nanog) indeed serve as pioneer factors, I profiled our nucleosome signal (based on MNase-seq data described in more detail in Chapter 4) around their binding sites, and only showing Nanog binding sites as an example. As expected, at the center of strong open sites and open sites (indicated by dark and light blue strip in 16A), nucleosome is depleted (Figure 3.16C, D). Strong open sites also exhibited even lower nucleosome signal compared to open sites. In contrast, closed sites were occupied by nucleosomes, confirming the *k*-mean clustering results. Nanog binding sites were called based on ChIP-seq of Nanog at 4.3hpf. If Nanog functions as pioneer factors on the closed sites, I should

observe lower nucleosome signals at post-MBT (5.3hpf) compared to MBT (4hpf). However, nucleosome signals at post-MBT were still high at the closed sites, which is contradictory to our hypothesis. More investigations are needed to resolve this contradiction.

3.3.10 ATAC-mono represents nucleosome positioning around transcription start sites

As I initially defined, reads with length from 120 to 220bp were grouped as ATAC-mono, which are close to a nucleosome size. To investigate whether ATAC-mono can be used to profile nucleosomes, I ranked all genes based on their expression at MBT, and then profiled ATAC-mono and my own MNase-seq data around TSS. ATAC-mono at MBT, post-MBT and liver displayed a banding pattern around TSS, which indicates nucleosome array. As expected, highly transcribed genes formed a stronger array compared to genes expressed at a low level (Figure 3.17A). In agreement with ATAC-mono data, MNase-seq data also revealed the presence of nucleosome array at TSS, and the strength of which correlates with gene expression (Figure 3.17B). Nevertheless, there are evident differences between these datasets. First, MNase-seq better represented nucleosomes distribution for gene transcribed at low level or nontranscribed genes (Figure 3.17B, bottom part of the heatmap). Second, stronger +2 and downstream nucleosomes were captured by MNase-seq data compared to ATAC-mono data. These differences were better illustrated in the class average maps, where ATAC-mono and MNase-seq data were compared directly at each stage (Figure 3.18). These differences likely stem from the nature of the two enzymes. MNase is a small enzyme ($\sim$17KD), whereas Tn5 is relative large ($\sim$55KD). In addition, MNase preferentially

cut at AT-rich sequence, which has high frequency between nucleosomes (linker regions). These attributes enable MNase to cut into heterochromatin regions, and also to have higher efficiency at linker regions. In conclusion, ATAC-mono can be used to reflect nucleosome positioning. However some information might be missing if solely relies on ATAC-mono.

3.4 Discussion

Zygotic genome activation (ZGA) is a crucial stage in early embryogenesis, where genes required for early development and transcription are selectively transcribed while genes involved in cell fate specification are repressed. How to tightly regulate this unique transcriptome and acquire the potential to differentiate into any cell lineage are important questions that remain elusive. We want to provide insights into these questions by profiling the genome-wide chromatin accessibility in the early zebrafish embryo. Chromatin accessibility, modulated by chromatin remodelers, controls transcription factors binding at promoters and enhancers, and the subsequent gene expression. Profiling the chromatin accessibility will allow us to understand how the unique transcriptome is regulated at the chromatin level and what the crucial regulators in early development are.

In my results, the number of open chromatin regions at pre-MBT is significantly lower compared to other stages. I argue it reflects a relative refractory chromatin state, which could be explained in three ways. First, the genome is transcriptionally quiescent. Thus, there is no need for the chromatin to be open. Second, there are repressors to prevent the genome from becoming accessible. Third, the factors responsible for opening up chromatin are not available or function at a low level.

These mechanisms could be acting individually or coordinately. However, it is also possible that the refractory chromatin state is a technical artifact, due to the condensed chromatin at mitosis. At pre-MBT, the cell cycle is very short ($\sim$15min), and only has S and M phase. Therefore, Tn5 enzyme may be inhibited by the compacted chromatin during M phase. However, there are no effective means to synchronize embryos at S phase. Nevertheless, the relative inaccessible genome is also observed in mouse preimplantation embryos by two different techniques, ATAC-seq and DNA hyperactivity assay (Lu et al., 2016; Wu et al., 2016). This makes the technical artifact hypothesis less likely, and supports that the less accessible chromatin in pre-MBT is biological.

A central question is how chromatin accessibility is established by the interplay between transcription factors and chromatin remodelers, which can be informed partly by the motif analysis on open chromatin regions. Many interesting motifs are over-represented in our accessible chromatin regions at different stages. At pre-MBT, motifs of Nr2e3, TBX and ZNF354C are present in almost 80% of open chromatin regions. They could serve as pioneer factors and open up the local chromatin at pre-MBT. Alternatively, they could bind these sites after the local chromatin becomes accessible. It will be equally interesting no matter which is the case. At MBT, post-MBT and liver, more candidates are suggested based on the identified motifs. Some factors including Pou5f1 and Sox have been shown to play important roles in ZGA. Others, including NFYB, CREB1, and YY1, have not been studied for their functions in early embryogenesis. Interestingly, a close homolog of NFYB, NFYA was recently identified in DHS (DNA hypersensitive sites) of mouse 2-cell embryos (Lu et al., 2016). The authors suggest NFYA may function

as a pioneer factor, as knockdown of NFYA reduces signals of ∼28% of DHSs (Lu et al., 2016). Furthermore, NFYA KD embryos are arrested at the morula stage. The evidence strongly suggests that zebrafish NFYB might have comparable roles in early embryogenesis.

Pluripotent factors are key regulators of early embryogenesis. It is interesting to determine whether they could serve as pioneer factors, in addtion to their roles as transcription activators. I have focused on three pluripotent factors Pou5f3, Sox2 and Nanog and one Tbx factor, Eomesa, which is homolog to Eomes in mammals that promotes trophectoderm differentiation. By *k*-mean clustering, two distinct groups of binding sites emerged for all these four factors. One group has a higher signal of ATAC-open data, suggesting open chromatin. The other group has much weaker ATAC-open signal, and the chromatin is probably inaccessible. Take Nanog as an example; I indeed observed a strong nucleosome peak right at the center of the closed sites, which are absent in the sites with open chromatin. Unexpectedly, the binding event of Nanog did not promote a subsequent decrease in nucleosome density. It is possible that the change of nucleosome density happens at later stages which are not covered in this dissertation work. Alternatively, the binding events have purposes other than increasing chromatin accessibility. In either case, further investigation is needed to determine whether these factors are bona fide pioneer factors. One essential role of transcription factors binding is to bring in co-activators such as histone acetyltransferases. By targeting those histone modifiers, active histone modifications such H3K27ac, H3K4me1/3 can be added, and further enhance the accessibility of these regions. Therefore, it will be interesting to determine whether the binding of these transcription factors affects

histone modifications.

In my results, I described a new bivalent mark at gene promoters, which is composed of H3K27me3 and H3K4me1. This new bivalent mark is different from the canonical one (H3K27me3 and H3K4me3). First, it will be interesting to know what genes carry the new bivalent marks and what biological processes they are enriched for. I would speculate genes such as Tbx, Fox or Hox genes, which are required for cell fate specification and need to be repressed at MBT, could carry this new bivalent histone modification. The repression would presumably be mediated through H3K27me3, while H3K4me1 may facilitate the genes to transit from a repressed to active state at proper cellular context. In ES cells, the active mark is H3K4me3 on the poised genes. Why do zebrafish utilize H3K4me1 instead of H3K4me3? This is not because H3K4me3 is limited, as actively transcribed genes are highly enriched for H3K4me3. It is possible that zebrafish employed a different strategy to distinguish the actively transcribed genes and poised genes. Therefore, additional questions to pursue are: 1) if H3K4me1 is replaced by H3K4me3 at their promoters once these bivalent genes are transcribed? 2) whether activation of genes is due to the removal of H3K27me3 or owing to establishment of H3K4me3? These questions can be answered by a combination of bioinformatics analysis, manipulation of histone methyltransferase, and ChIP-seq experiments.

Previously, I have defined Open Pol III genes at each stage by their association with open chromatin around the TSS. Open chromatin at promoter indicates their potential for transcription but does not necessarily mean they are transcribed. Here, Pol III ChIP-seq will help us validate whether these Pol III genes are transcribed, which will be investigated by Candice Wike in the lab. Pol III ChIP data

are also helpful to confirm the dynamics of Pol III during early embryogenesis. By combining her data with my data I could address if my new bivalent modifications also mark a subset of Pol III genes that are poised for transcription. In addition, transcription of Pol I genes might also be regulated at chromatin level (Parnell et al. 2008). Similar analysis for Pol III genes can be applied to Pol I to address this question. However, due to the repetitiveness of the rRNA, reads need to be mapped to rRNA locus using different alignment strategy.

In Chapter 3, I have described the ATAC-seq data generated from early embryos and adult liver, with a focus on ATAC-open data (fragments with length from 0-100bp). Our data revealed the accessibility of chromatin increases dramatically from pre-MBT to MBT, which suggests a permissive chromatin is underlining the genome activation. This work also provides candidates transcription factors that are seminal for the establishment of chromatin accessibility and in the regulation of early embryogenesis. Furthermore, I have also observed various patterns of histone modifications at promoters and intergenic regions. Particularly, a new bivalent mark composed of H3K4me1 and H3K27me3 is identified at promoters of certain genes. Future work should focus on studying the candidate transcription factors that are potentially involved in early embryogenesis, and characterization of the bivalent histone mark and further investigating its function.

3.5 Methods

3.5.1 Zebrafish and sample collection

Tubingen zebrafish lines were maintained and raised under standard conditions. Wildtype embryos were collected after 10min of mating to ensure the syn-

chrony of embryo development, raised in embryo water at 28°C and staged as previously described (Kimmel et al., 1995).

3.5.2 Isolation of adult liver

Adult fish were fasted for a week prior to dissection. Euthanasia of zebrafish with ice water was conducted following the IACUC procedures. Dissection of the adult zebrafish liver was performed under brightfield imaging on a dissection microscope. Dissected livers were immediately transferred to Eppendorf tubes on ice and rinsed with PBS. Nuclei extraction is similar to embryonic cells, and is described in the following section.

3.5.3 ATAC-seq experiment

ATAC-seq was performed essentially as previously described (Buenrostro et al., 2013). Briefly, embryos at pre-MBT (2.5hpf), MBT (4hpf) and post-MBT (5.3hpf) were carefully staged, and dechorinated. The embryos were washed in cold PBS, and then resuspended in lysis buffer (10mM Tris-HCl [pH 7.4], 10mM NaCl, 3mM MgCl2, 0.1% IGEPAL CA-630) on ice for 10 min. During the lysis, a 20 Gauge syringe was used three times to break the chorions and cells, followed by table top spin for 10 seconds. The supernatant was then centrifuged at 1300g 5 min 4°C and nuclei was resuspended in 2X TD buffer, and incubated with Tn5 enzyme for 30 min at 55°C (Nextera DNA Library Preparation Kit, FC-121-1031). The reaction was terminated by adding SDS to a final concentration 1%. Then, the sample was purified by Qiagen minElute column and PCR amplified by NEBNext High-Fidelity 2X PCR Master Mix (NEB, M0541L). qPCR was utilized to determine the optimal PCR cycles to prevent over-amplification.

3.5.4 Data analysis

Paired-end fastq files were uniquely aligned to zv10 zebrafish genome assembly using Novoalign (Novocraft, Inc) with the following parameters: `-r None -Q 13 -k -o SAM -a CTGTCTCTTATACACATCT`, and converted to BAM files using Ssamtools. BAM files from replicates were merged for downstream analysis. Total ATAC-seq data are divided into nucleosome-free reads and nucleosome signal according to (Buenrostro et al., 2013). The coverage tracks of nucleosome-free reads were generated using the program bam2wig (`http://search.cpan.org/dist/Bio-ToolBox/`) with the following parameters: `--position span --pe --rpm --bw`.

3.6 Contribution and acknowledgement

This work in combination with the work in Chapter 4 are under preparation for publication. The ATAC-seq experiments were performed collaboratively with Candice Wike. I have performed all the bioinformatic analysis. I want to thank Tim Parnell and Tim Mosbruger for bioinformatics assistance, and Brian Dalley for sequencing expertise.

3.7 References

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bogdanović, O., Fernandez-Miñán, A., Tena, J.J., de la Calle-Mustienes, E., Hidalgo, C., van Kruysbergen, I., van Heeringen, S.J., Veenstra, G.J.C., Gómez-Skarmeta, J.L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Research *22*, 2043–2053.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open

chromatin, dna-binding proteins and nucleosome position. Nature methods *10*, 1213–1218.

Burton, A., Torres-Padilla, M.E. (2014). Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. Nature Reviews Molecular Cell Biology *15*, 723–735.

Hartley, P.D., Madhani, H.D. (2009). Mechanisms that specify promoter nucleosome location and identity. Cell *137*, 445–458.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Developmental Dynamics *203*, 253–310.

Leichsenring, M., Maes, J., Mössner, R., Driever, W., Onichtchouk, D. (2013). Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science *341*, 1005–1009.

Levine, M., Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147–151.

Lu, F., Liu, Y., Inoue, A., Suzuki, T., Zhao, K., Zhang, Y. (2016). Establishing chromatin regulatory landscape during mouse preimplantation development. Cell *165*, 1375–1388.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., et al. (2015). Rsat 2015: regulatory sequence analysis tools. Nucleic Acids Research , gkv362.

Nelson, A.C., Cutty, S.J., Niini, M., Stemple, D.L., Flicek, P., Houart, C., Bruce, A.E., Wardle, F.C. (2014). Global identification of smad2 and eomesodermin targets in zebrafish identifies a conserved transcriptional network in mesendoderm and a novel role for eomesodermin in repression of ectodermal gene expression. BMC Biology *12*, 1.

Niwa, H. (2007). How is pluripotency determined and maintained? Development *134*, 635–646.

Oler, A.J., Alla, R.K., Roberts, D.N., Wong, A., Hollenhorst, P.C., Chandler, K.J., Cassiday, P.A., Nelson, C.A., Hagedorn, C.H., Graves, B.J., et al. (2010). Human rna polymerase iii transcriptomes and relationships to pol ii promoter chromatin and enhancer-binding factors. Nature Structural & Molecular Biology *17*, 620–628.

Parnell, T.J., Huff, J.T., Cairns, B.R. (2008). Rsc regulates nucleosome positioning at pol ii genes and density at pol iii genes. The EMBO Journal *27*, 100–110.

Russ, A.P., Wattler, S., Colledge, W.H., Aparicio, S.A., Carlton, M.B., Pearce, J.J., Barton, S.C., Surani, M.A., Ryan, K., Nehls, M.C., et al. (2000). Eomesodermin is required for mouse trophoblast development and mesoderm formation. Nature *404*, 95–99.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., et al. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature *534*, 652–657.

Xu, C., Fan, Z.P., Müller, P., Fogley, R., DiBiase, A., Trompouki, E., Unternaehrer, J., Xiong, F., Torregroza, I., Evans, T., et al. (2012). Nanog-like regulates endoderm formation through the mxtx2-nodal pathway. Developmental Cell *22*, 625–638.

Zaret, K.S., Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes & Development *25*, 2227–2241.

Zhang, Y., Vastenhouw, N.L., Feng, J., Fu, K., Wang, C., Ge, Y., Pauli, A., van Hummelen, P., Schier, A.F., Liu, X.S. (2014). Canonical nucleosome organization at promoters forms during genome activation. Genome Research *24*, 260–266.

Table 3.1: ATAC-seq sequencing summary

	Pairs	**Pairs aligned**	**% Aligned**	**Correlation**
pre-MBT_nCL_rep1	127,360,630	108,799,541	85.4%	0.9997
pre-MBT_nCL_rep2	155,469,925	133,997,672	86.2%	
MBT_CL_rep1	141,697,661	113,722,176	80.3%	0.9913
MBT_CL_rep2	88,064,503	70,400,750	79.9%	
MBT_nCL_rep1	116,143,770	100,727,211	86.7%	0.9970
MBT_nCL_rep2	69,279,042	94,374,321	136.2%	
post-MBT_nCL_rep1	144,010,745	126,396,904	87.8%	0.9989
post-MBT_nCL_rep2	104,215,950	90,975,359	87.3%	
Liver_nCL_rep1	102,588,621	85,385,785	83.2%	0.9914
Liver_nCL_rep2	146,966,785	124,930,294	85.0%	
Naked_DNA	228,249,656	208,580,057	91.4%	

Table 3.2: Pol III genes

Pol III genes	5s rRNA	7SK RNA	vault RNA	Y RNA	U6 snRNA	tRNA
Numbers	758	2	4	1	706	12465

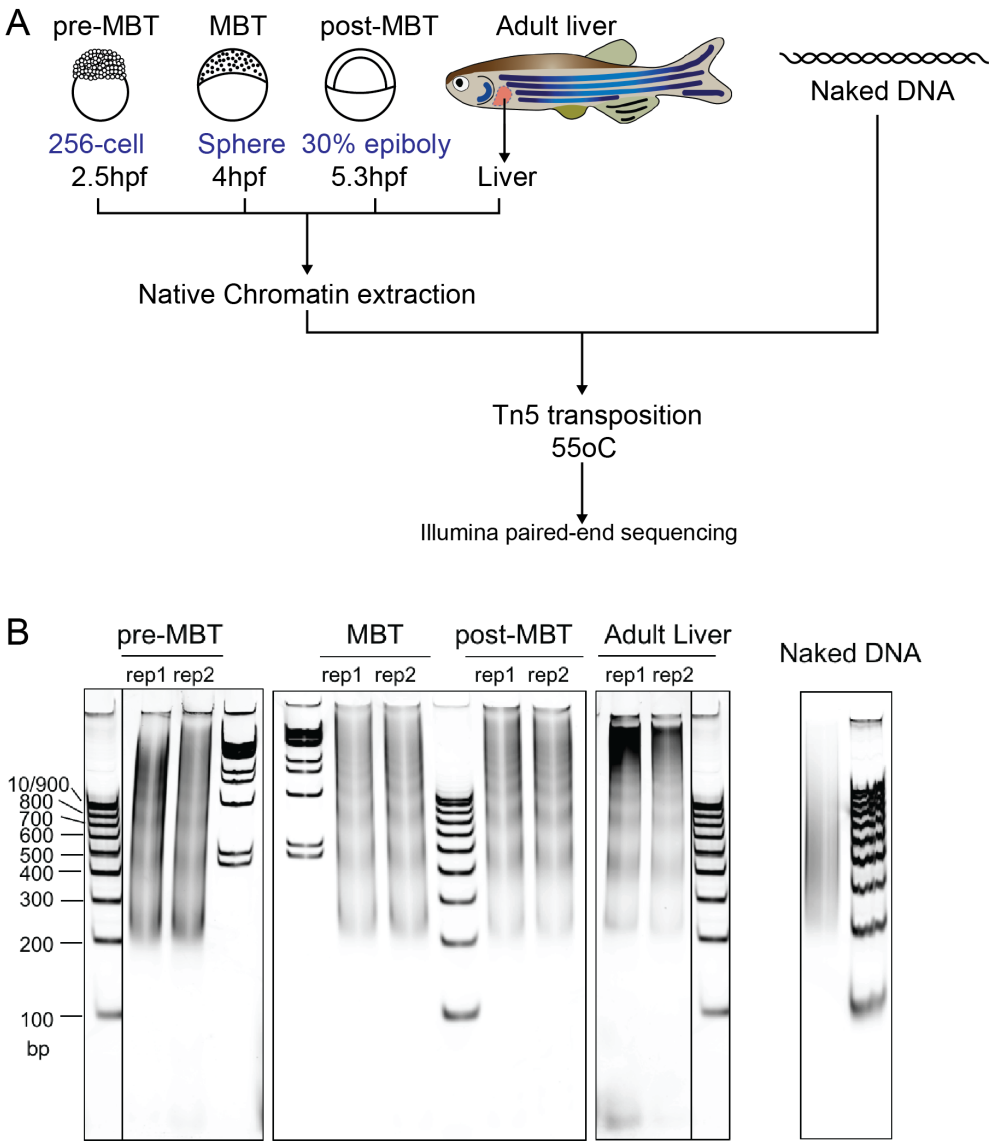
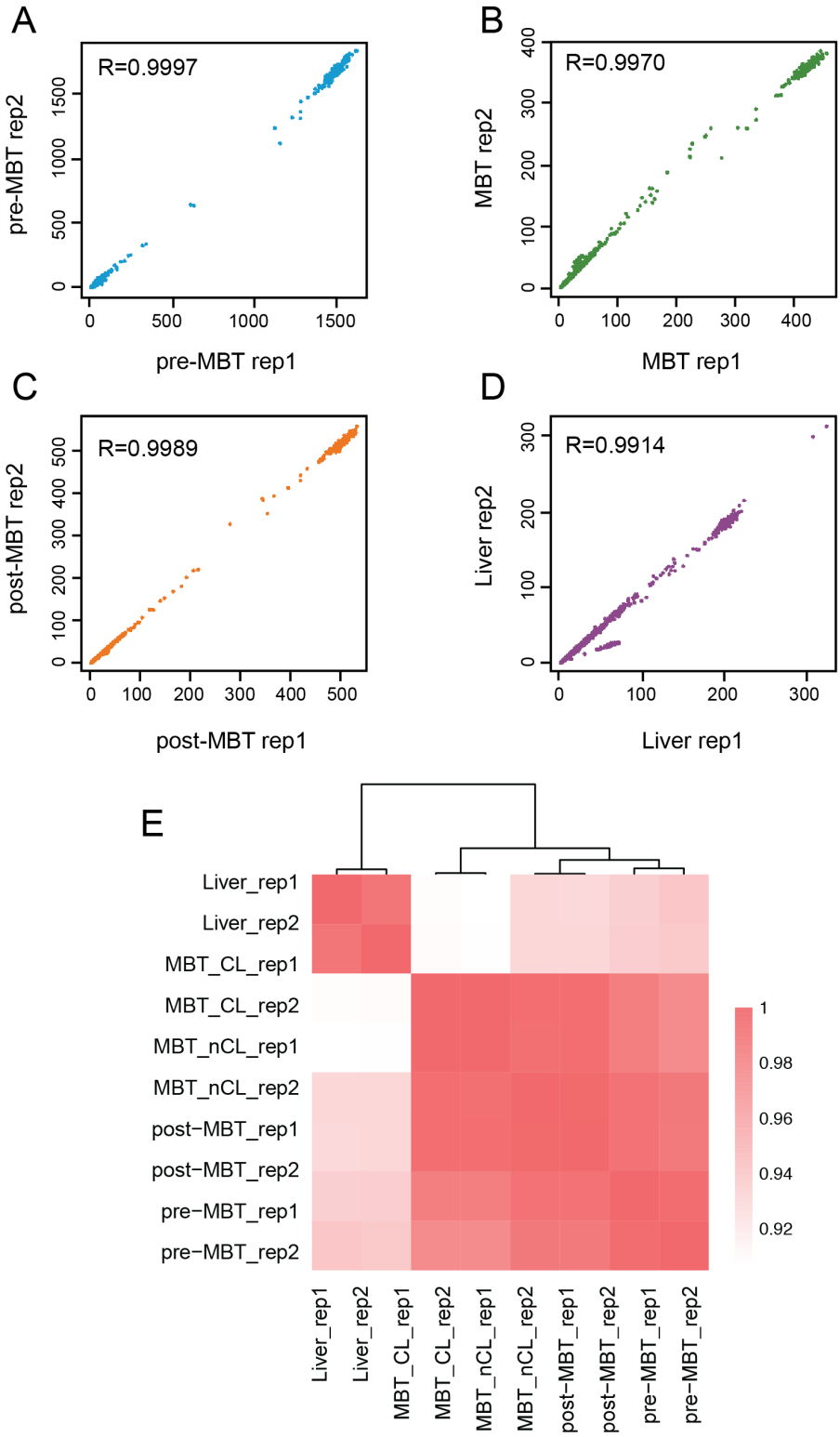


Figure 3.1: Overview of ATAC-seq experiment. (A) Schematic of experiment design. Native chromatins are extracted from pre-MBT, MBT, post-MBT embryos, and adult liver, and are subject to Tn5 transposition and paired-end sequencing. Naked DNA is treated as chromatin samples and serves as a control. (B) TBE gels of DNA fragments generated by transposition experiment of two replicates at each stage.

Figure 3.2: Replicates are significantly correlated. (A) Pairwise comparisons of replicates at pre-MBT (A), MBT (B), post-MBT (C) and liver (D). For each replicate, ATAC signals are collected at 100bp windows covering the entire genome. R value represents Pearson Correlation score. (E) Pearson Correlations between any of the 8 samples, clustered into heatmap.


A
R=0.9997
pre-MBT rep2
pre-MBT rep1
0
500
1000
1500
B
R=0.9970
MBT rep2
MBT rep1
0
100
200
300
400
C
R=0.9989
post-MBT rep2
post-MBT rep1
0
100
200
300
400
500
D
R=0.9914
Liver rep2
Liver rep1
0
100
200
300
E
Liver_rep1
Liver_rep2
MBT_CL_rep1
MBT_CL_rep2
MBT_nCL_rep1
MBT_nCL_rep2
post−MBT_rep1
post−MBT_rep2
pre−MBT_rep1
pre−MBT_rep2
1
0.98
0.96
0.94
0.92

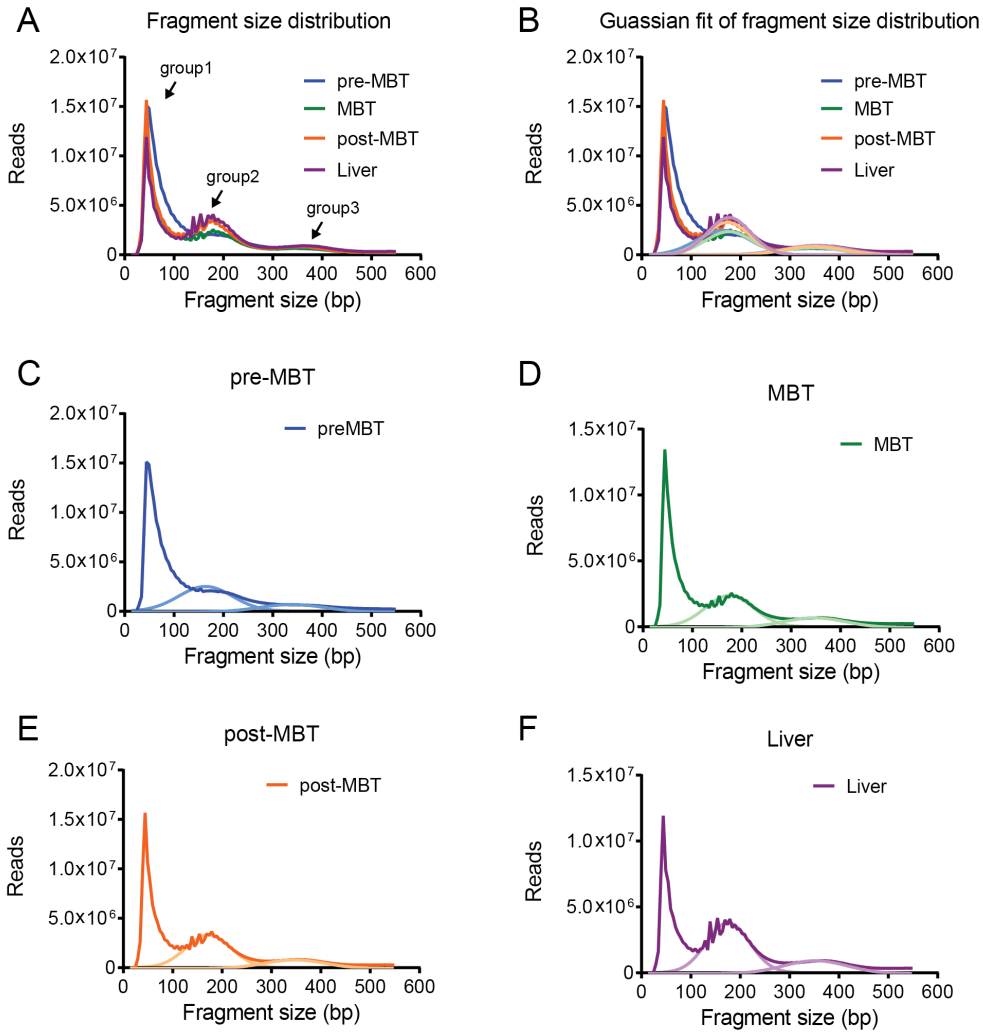


Figure 3.3: Fragment size distribution. (A) Fragment size distribution for merged samples at pre-MBT, MBT, post-MBT and liver. Arrows indicate the three groups of reads. (B) Fragment size distributions of all samples are fitted with Gaussian distribution. (C-F) Fragment size distributions of each stage were fitted with Gaussian distribution.

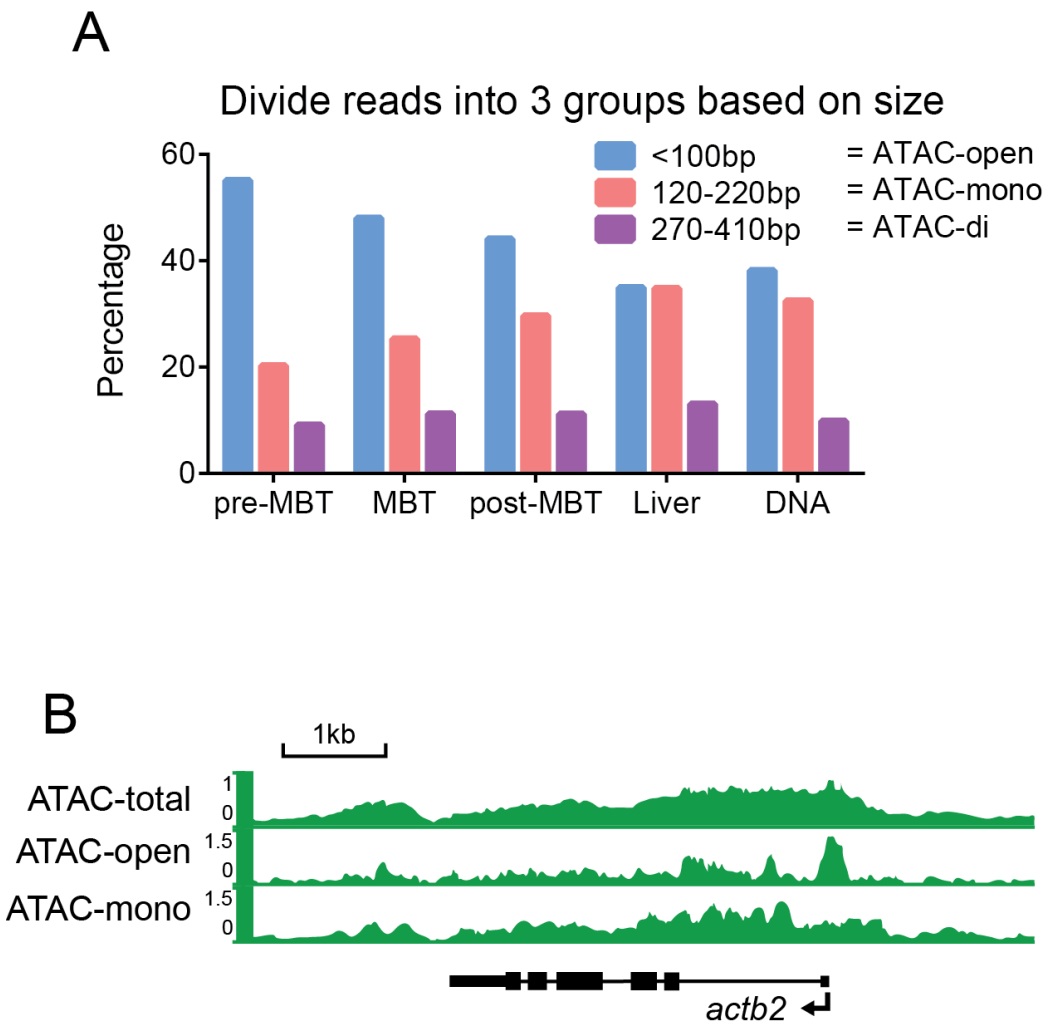


Figure 3.4: ATAC-seq reads are separated into 3 groups based on size. (A) Percentage of each group at all stages. (B) Genome browser snapshots of ATAC-total, ATAC-open and ATAC-mono.

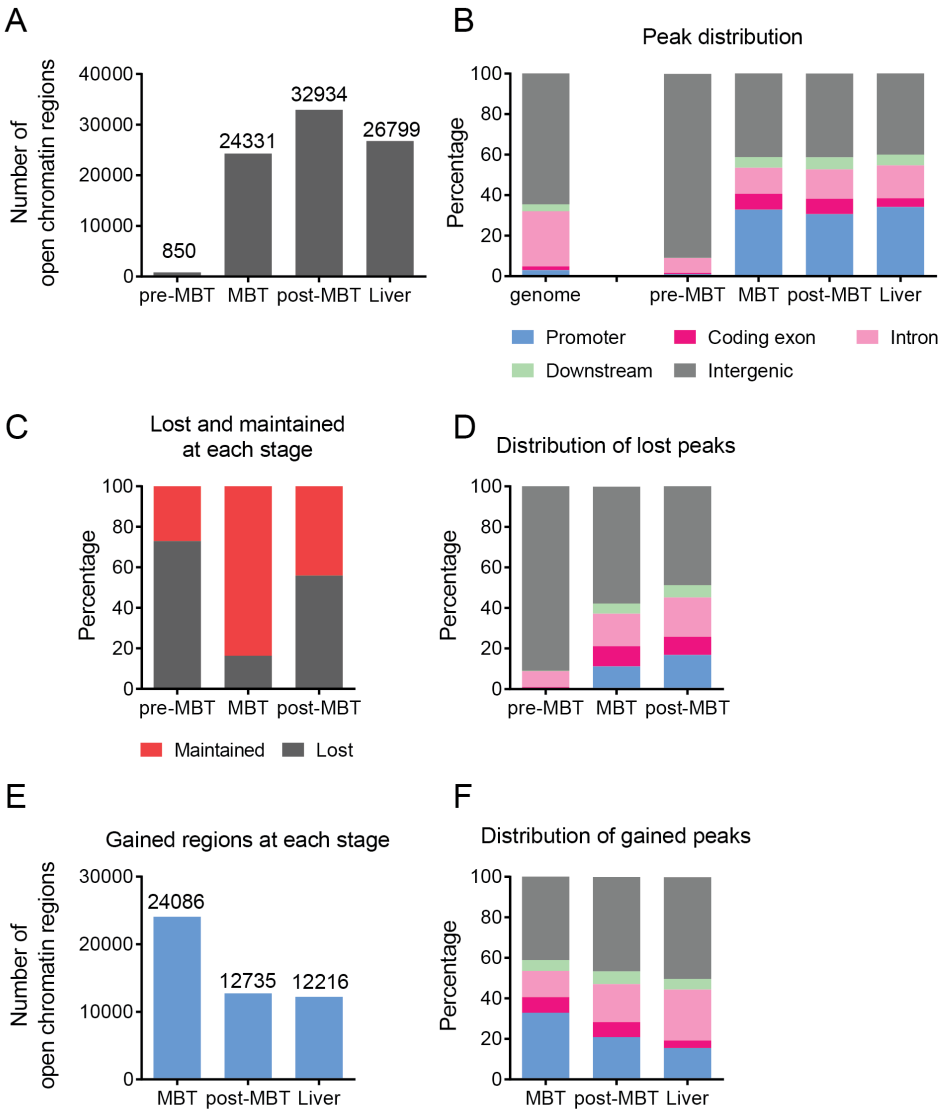
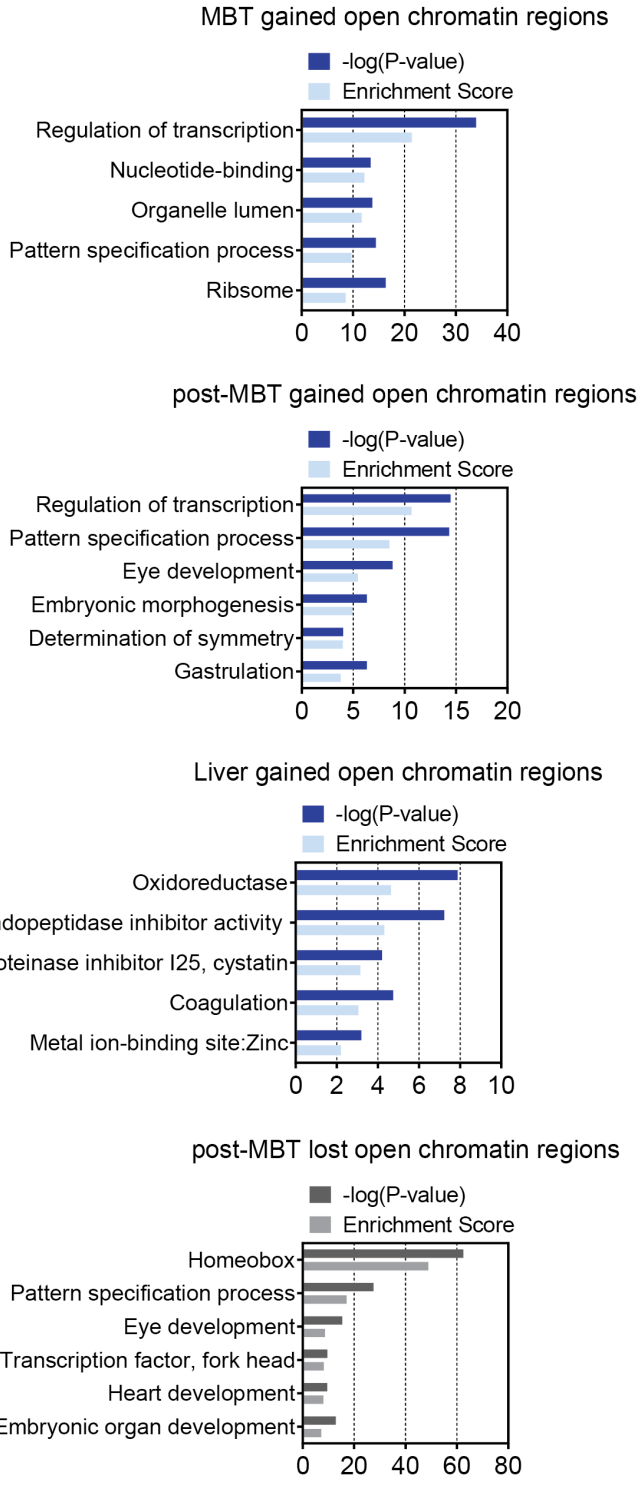


Figure 3.5: Dynamics of open chromatin regions. (A) Numbers of ATAC-open peaks called by MACS2 and R filtration at each stage. (B) Distribution of ATAC-open peaks at each stage on RefSeq annotation elements. (C) Percentage of lost and maintained peaks at each stage comparing to the subsequent stage. (D) same as in (B) for ATAC-open peaks that are lost in the subsequent stage. (E) Number of ATAC-open peaks that are gained at each stage. (E) same as in (B) for ATAC-open peaks that are gained comparing to its previous stage.

Figure 3.6: Gene ontology of lost and gained ATAC-open peaks. Gene ontology of peaks gained at MBT, post-MBT and liver, as well as post-MBT peaks that are lost in liver. Analysis is performed with the online server, DAVID.


MBT gained open chromatin regions
-log(P-value)
Enrichment Score
Regulation of transcription
Nucleotide-binding
Organelle lumen
Pattern specification process
Ribsome
0 10 20 30 40
post-MBT gained open chromatin regions
-log(P-value)
Enrichment Score
Regulation of transcription
Pattern specification process
Eye development
Embryonic morphogenesis
Determination of symmetry
Gastrulation
0 5 10 15 20
Liver gained open chromatin regions
-log(P-value)
Enrichment Score
Oxidoreductase
Endopeptidase inhibitor activity
Proteinase inhibitor I25, cystatin
Coagulation
Metal ion-binding site:Zinc
0 2 4 6 8 10
post-MBT lost open chromatin regions
-log(P-value)
Enrichment Score
Homeobox
Pattern specification process
Eye development
Transcription factor, fork head
Heart development
Embryonic organ development
0 20 40 60 80

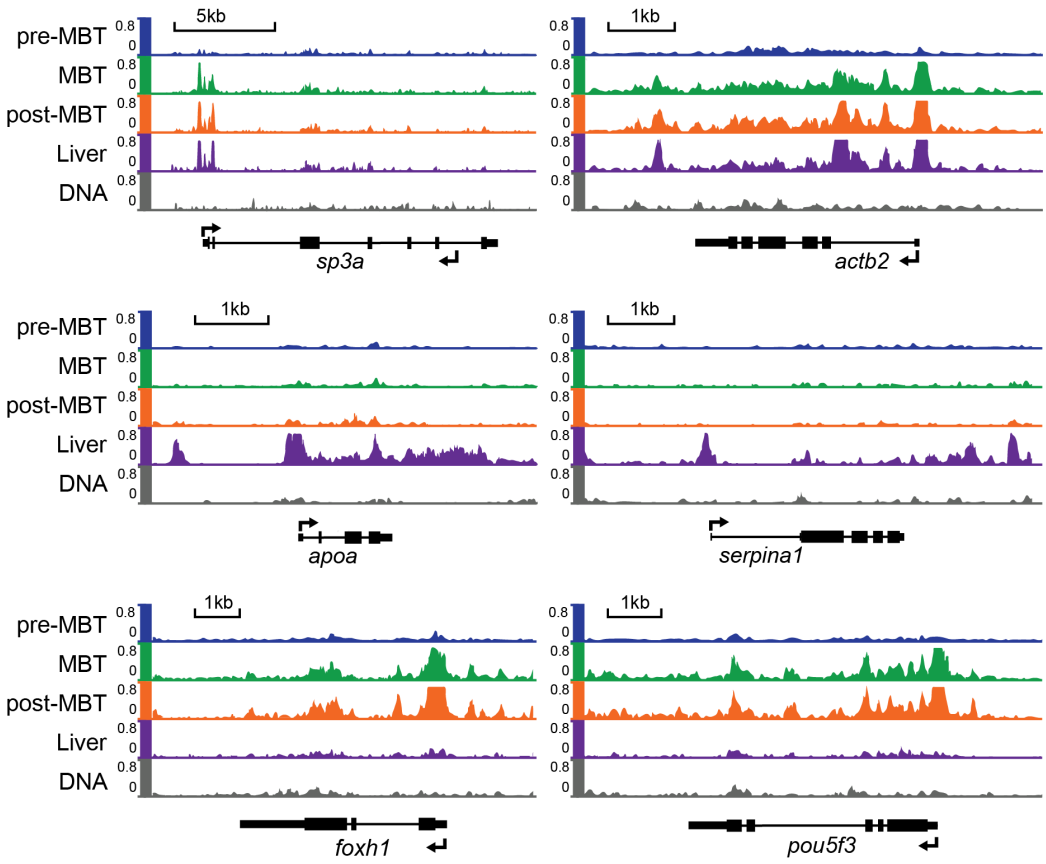


Figure 3.7: Genome browser snapshots of representative loci. Sp3a and actb2 have ATAC-open signal at MBT, post-MBT and liver. apoa and serpina1 have ATAC-open signal specifically in liver. foxh1 and pou5f3 have ATAC-open signal only at MBT and post-MBT.

Figure 3.8: Genes associated with open chromatin have higher expression. (A) Numbers of genes associated with open chromatin at each stage. (B) Meta-gene model of mean profiles of ATAC-open signals on their occupied genes at each stage. (C) Boxplot of expression level of genes associated or not associated with open chromatin (t test: *** denotes $p<0.001$).

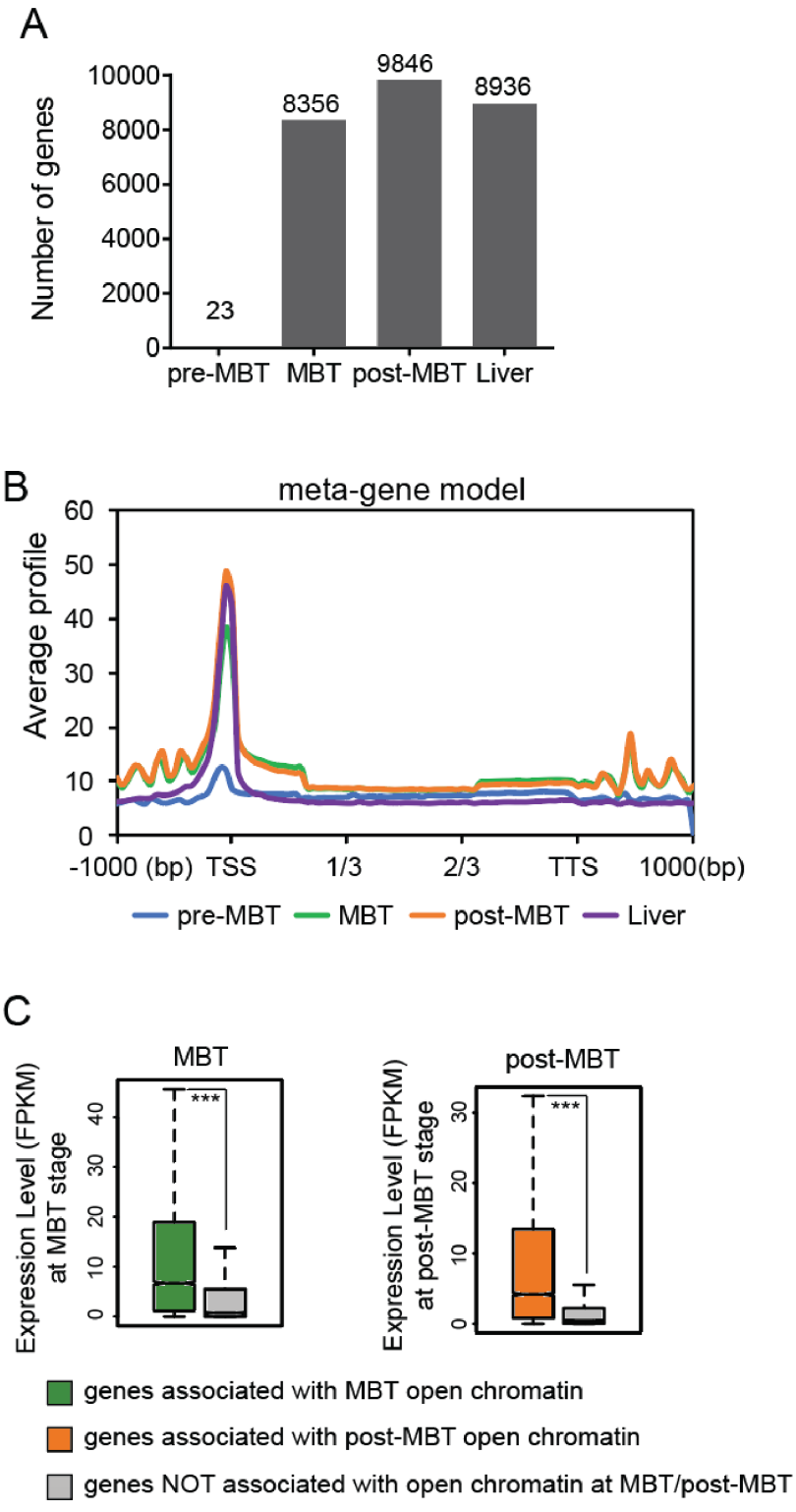
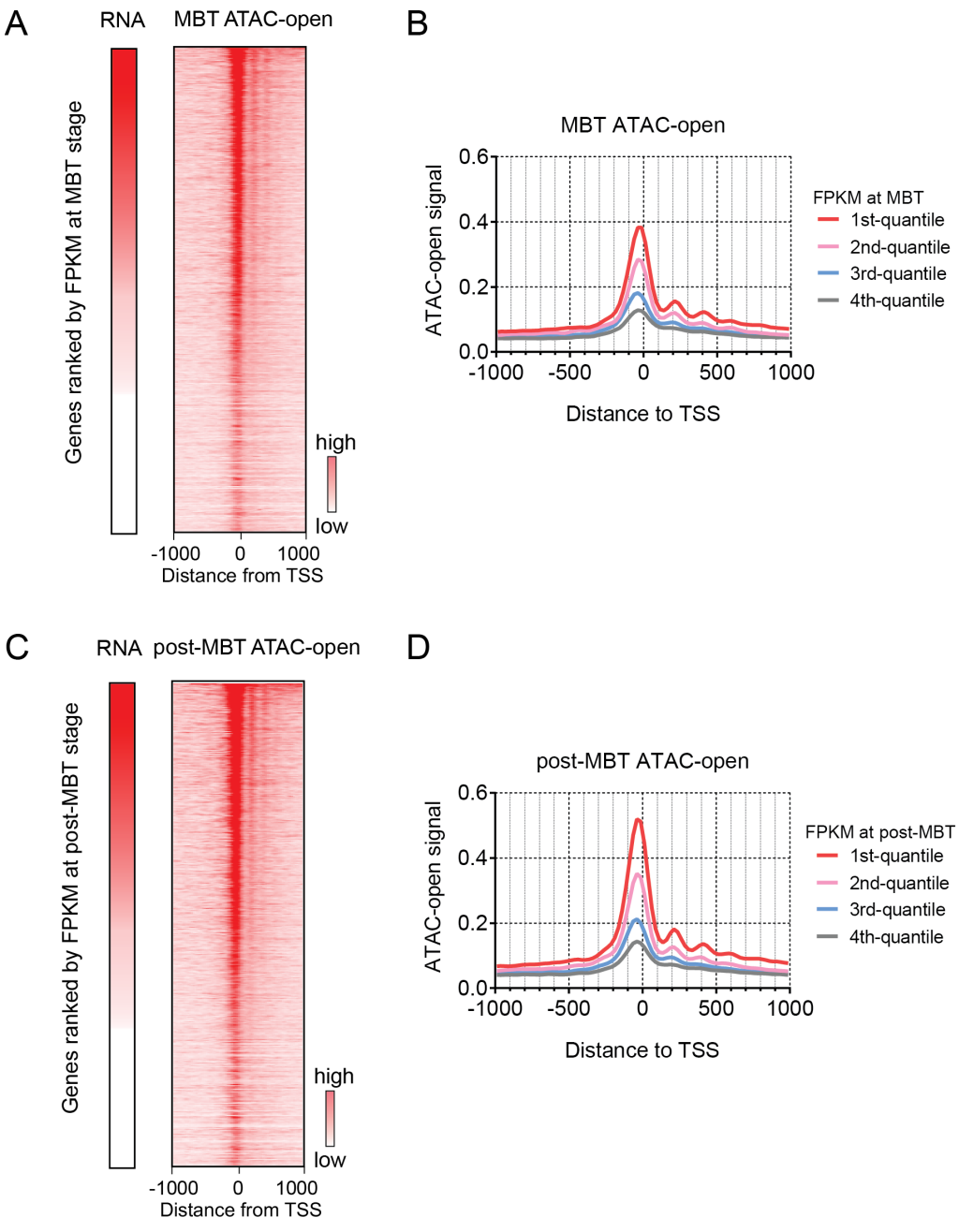

A
Number of genes
10000
8000
6000
4000
2000
0
23
8356
9846
8936
pre-MBT
MBT
post-MBT
Liver
B
meta-gene model
Average profile
60
50
40
30
20
10
0
-1000 (bp)
TSS
1/3
2/3
TTS
1000(bp)
pre-MBT
MBT
post-MBT
Liver
C
MBT
post-MBT
Expression Level (FPKM) at MBT stage
Expression Level (FPKM) at post-MBT stage
***
genes associated with MBT open chromatin
genes associated with post-MBT open chromatin
genes NOT associated with open chromatin at MBT/post-MBT

Figure 3.9: Highly transcribed genes have stronger open chromatin. (A) Heatmaps of MBT ATAC-open signal around TSS (1kb). Each row represents a Refseq gene TSS, ranked by RNA FPKM value at MBT (4hpf). (B) The mean profiles of MBT ATAC-open signals at promoters of 4 groups of genes. Genes are grouped into 4 quantiles based on the expression level (FPKM) at MBT. 1st quantile: top 25% expressed genes. 4th quantile: bottom 25% expressed genes. (C) similar as in (A) for post-MBT ATAC-open signal on genes ranked by expression at post-MBT. (D) similar as in (B), for post-MBT ATAC-open signals at promoters of 4 groups of genes grouped by expression at post-MBT.


A
RNA
MBT ATAC-open
Genes ranked by FPKM at MBT stage
high
low
-1000
0
1000
Distance from TSS
B
MBT ATAC-open
ATAC-open signal
0.6
0.4
0.2
0.0
-1000
-500
0
500
1000
Distance to TSS
FPKM at MBT
1st-quantile
2nd-quantile
3rd-quantile
4th-quantile
C
RNA
post-MBT ATAC-open
Genes ranked by FPKM at post-MBT stage
high
low
-1000
0
1000
Distance from TSS
D
post-MBT ATAC-open
ATAC-open signal
0.6
0.4
0.2
0.0
-1000
-500
0
500
1000
Distance to TSS
FPKM at post-MBT
1st-quantile
2nd-quantile
3rd-quantile
4th-quantile

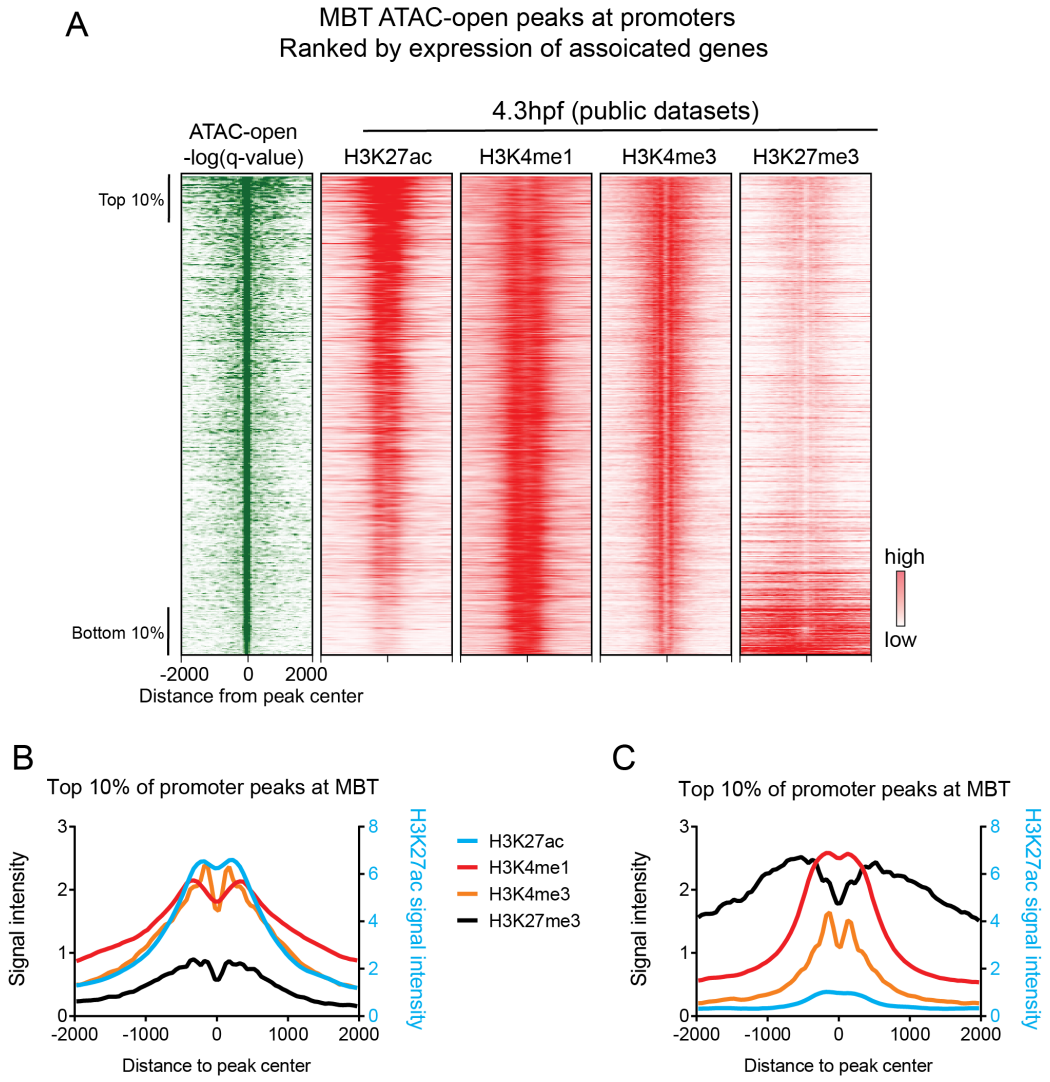


Figure 3.10: Bivalent genes have open promoters but are not transcribed. (A) Heatmaps of H3K27ac, H3K4me1, H3K4me3 and H3K27me3 (ChIP-seq signals) around MBT ATAC-open peaks residing at promoters. Green heatmap represents the mean $-$log(q-value) around the MBT ATAC-open peaks ($\pm$1kb). MBT ATAC-open peaks are ranked by expression level of associated genes. (B) The mean profiles of histone modifications signal around the top 10% of MBT ATAC-open peaks. (C) Same as in (B), for the bottom 10% of MBT ATAC-open peaks. Signals of H3K4me1, H3K4me3 and H3K27me3 are plotted on the left Y axis. Signal of H3K27ac is plotted on the right Y axis.

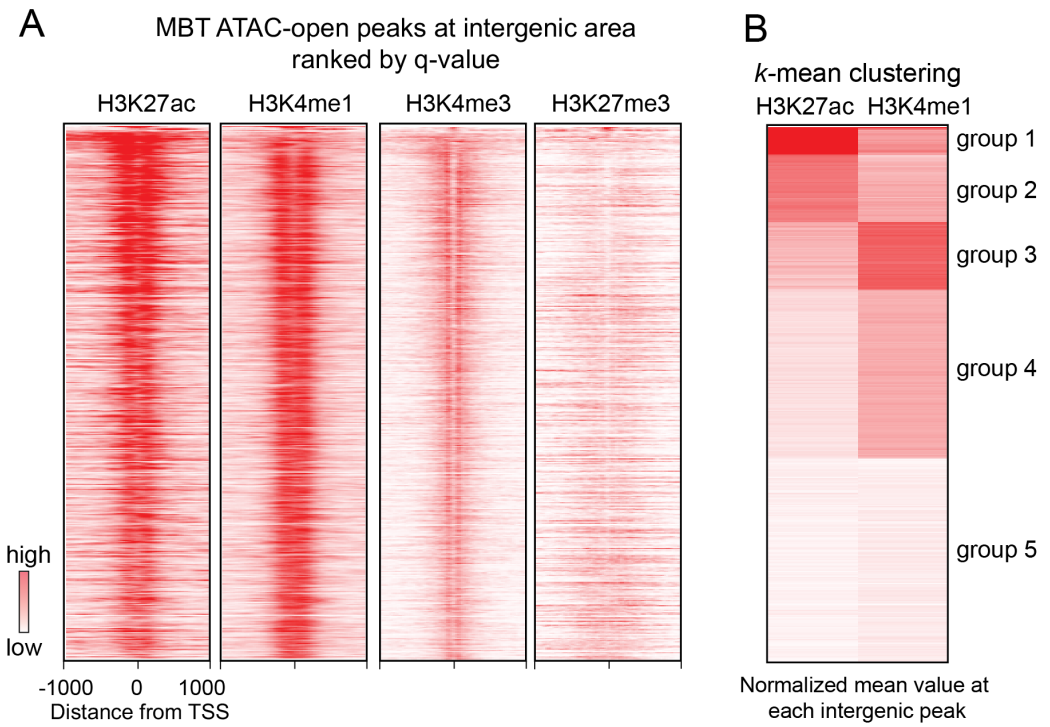


Figure 3.11: Intergenic ATAC-open peaks are enriched for H3K27ac and H3K4me1. (A) Heatmaps of H3K27ac, H3K4me1, H3K4me3 and H3K27me3 (ChIP-seq signals) around intergenic MBT ATAC-open peak ($\pm$2kb around peak center). (B) *k*-mean clustering of H3K27ac and H3K4me1 signals at intergenic MBT ATAC-open peaks. For each peak, mean values of H3K27ac and H3K4m1 were collected and normalized.

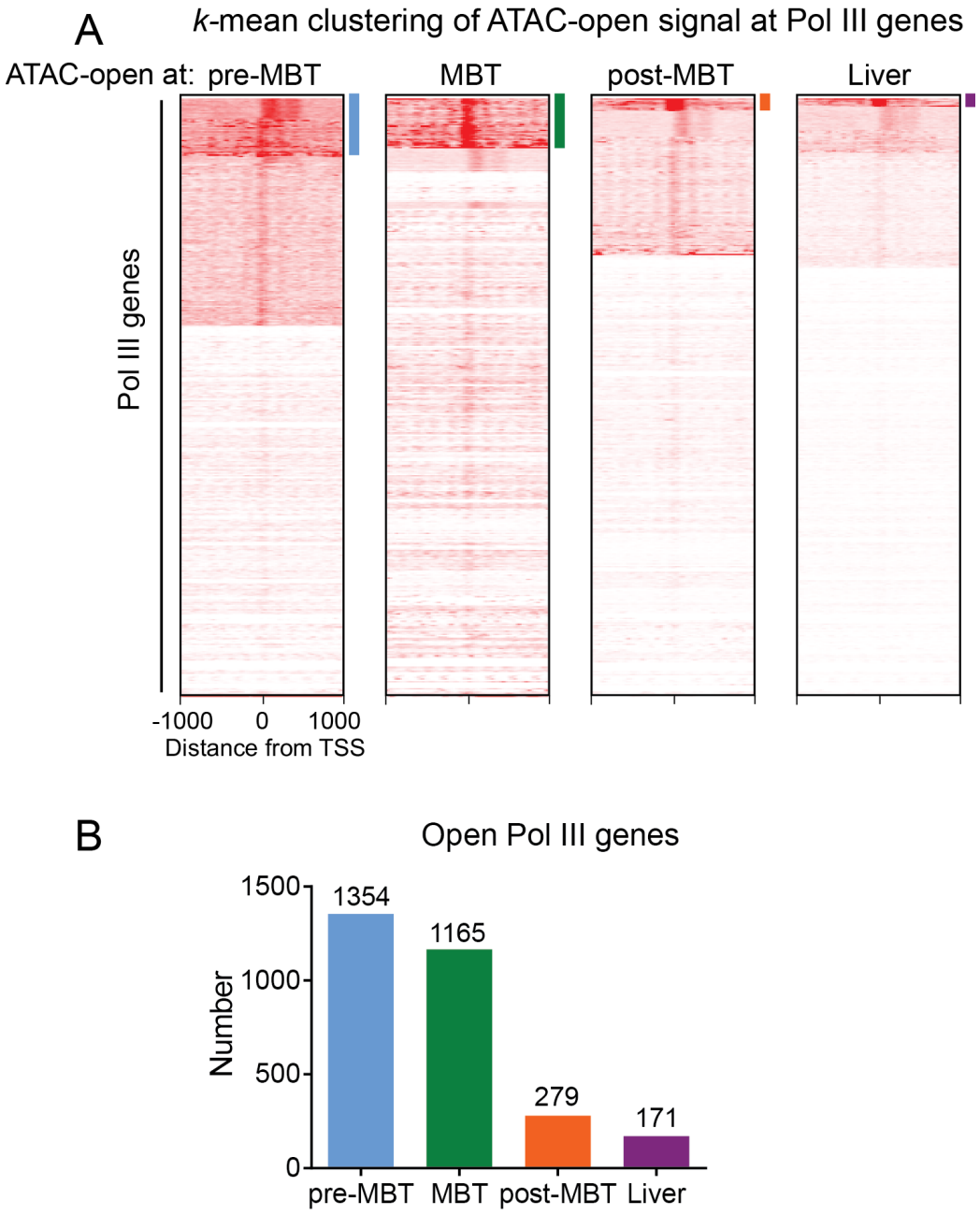


Figure 3.12: Pol III genes are relatively open at early stages. (A) *k*-mean clustering of ATAC-open signal at pre-MBT, MBT post-MBT, and liver around the TSS of all Pol III genes. The potential open Pol III genes are marked by color bar. (B) Number of open Pol III genes at each stage.

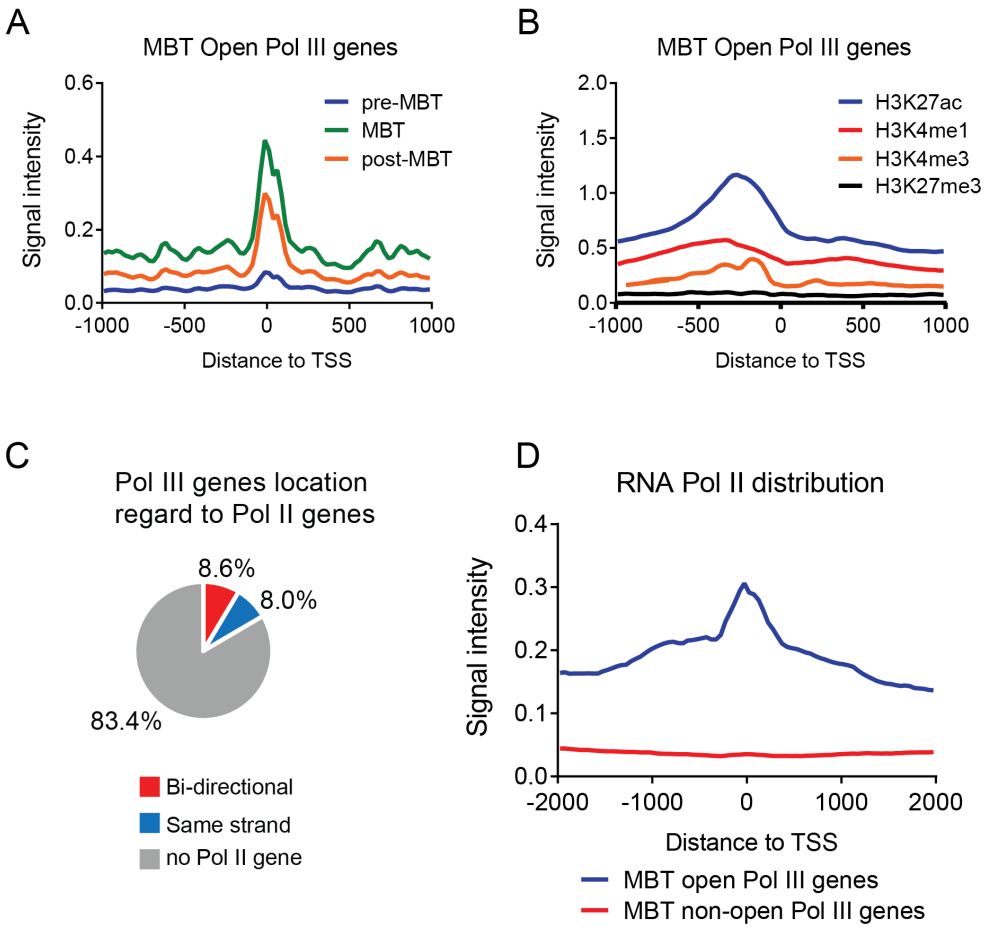


Figure 3.13: Correlation of Pol III genes with open chromatin, histone modifications and Pol II distribution. (A) The mean profiles of ATAC-open signal at pre-MBT, MBT and post-MBT around MBT open Pol III genes. (B) Same as in (A), for H3K27ac, H3K4me1, H3K4me3 and H3K27me3 (ChIP-seq signals). (C) Pie chart of the percentage of Pol III within 5kb of Pol II genes. (D) The mean profiles of Pol II signals (ChIP-seq at MBT) around MBT open Pol III genes and MBT non-open Pol III genes.

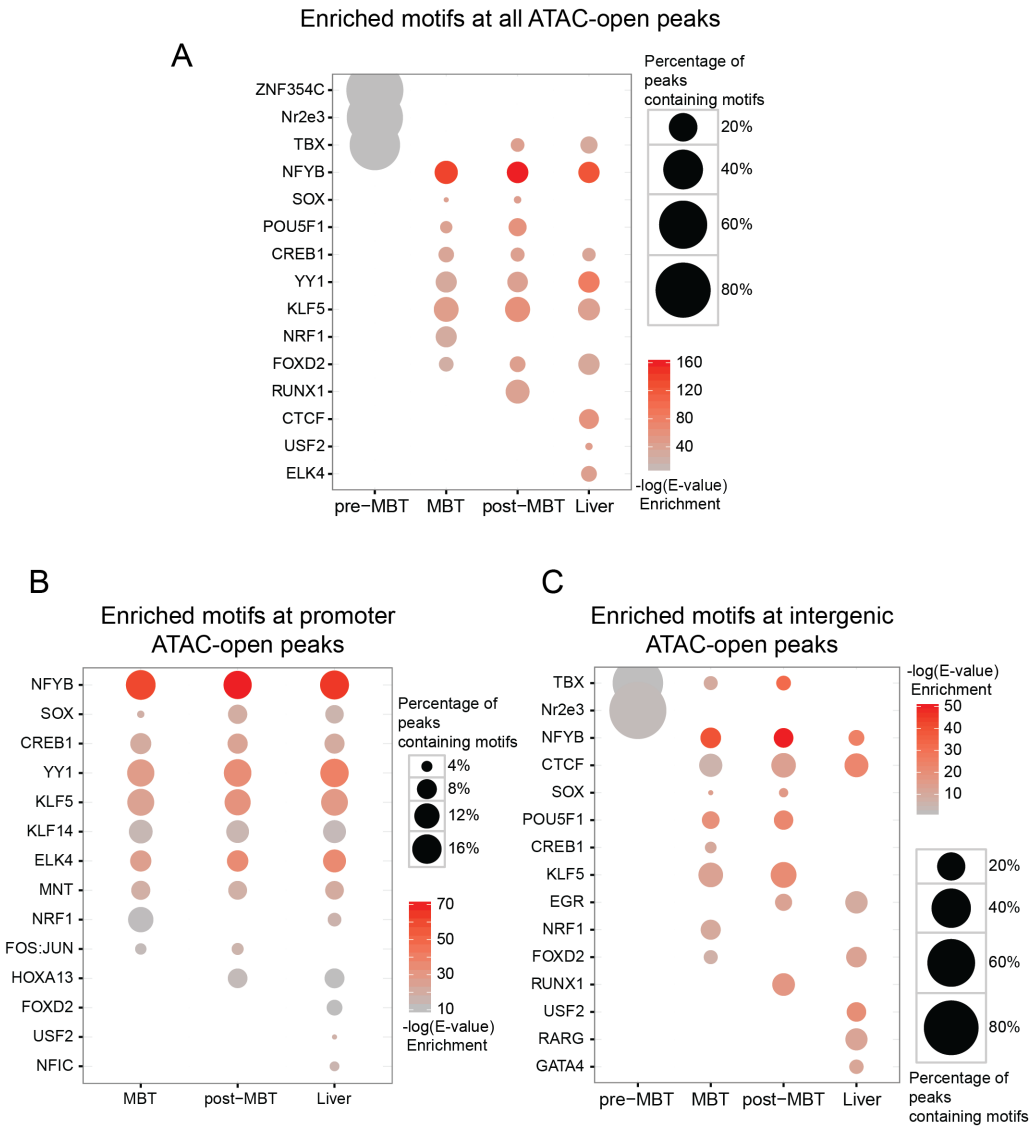
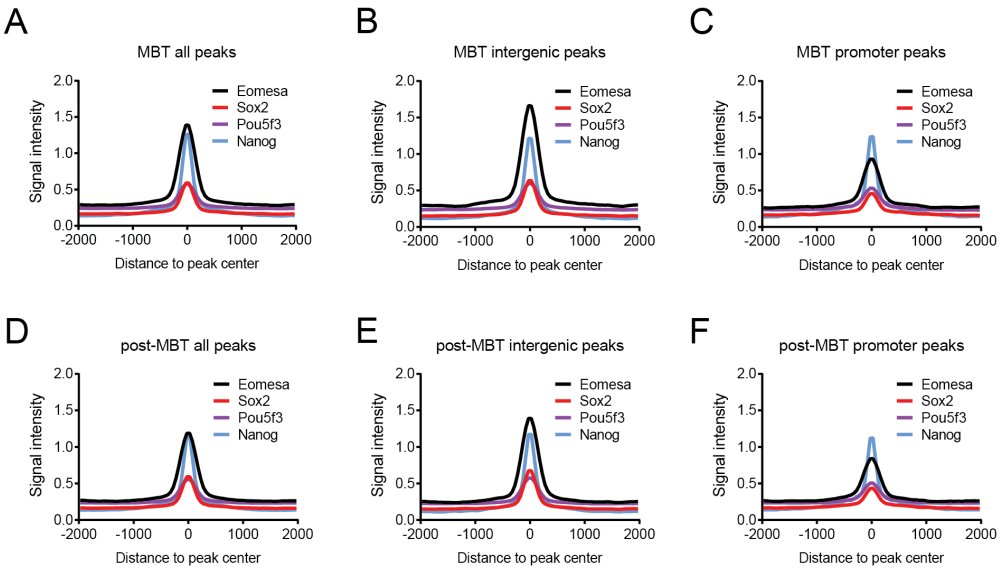


Figure 3.14: Over-represented motifs identified at ATAC-open peaks. (A) Top 10 enriched motifs reported by RSAT are summarized for all ATAC-open peaks at pre-MBT, MBT, post-MBT and liver. (B) Same as in (A), for promoter peaks at MBT, post-MBT and liver. (C) Same as in (A), for intergenic peaks at four stages (C).

Figure 3.15: Eomesa, Sox2, Pou5f3 and Nanog are enriched at open chromatin regions. (A-C) The mean profiles of Eomesa, Sox2, Pou5f3, and Nanog signals around MBT ATAC-open peaks ($\pm$2kb from the peak center). (A) all peaks, (B) intergenic peaks, (C) promoter peaks. (D-E) same as (A-C), for post-MBT ATAC-open peaks. Profiles of Eomesa, Sox2, Pou5f3 and Nanog are based on public ChIP-seq data from 4hpf, 5hpf, 5hpf and 4.3hpf embryos, respectively. (G, H) *k*-mean clustering of TFs signal at intergenic MBT ATAC-open peaks (G) and at promoter MBT ATAC-open peaks (H).


A
MBT all peaks
B
MBT intergenic peaks
C
MBT promoter peaks
D
post-MBT all peaks
E
post-MBT intergenic peaks
F
post-MBT promoter peaks
Eomesa
Sox2
Pou5f3
Nanog
Signal intensity
Distance to peak center
2.0
1.5
1.0
0.5
0.0
-2000
-1000
0
1000
2000

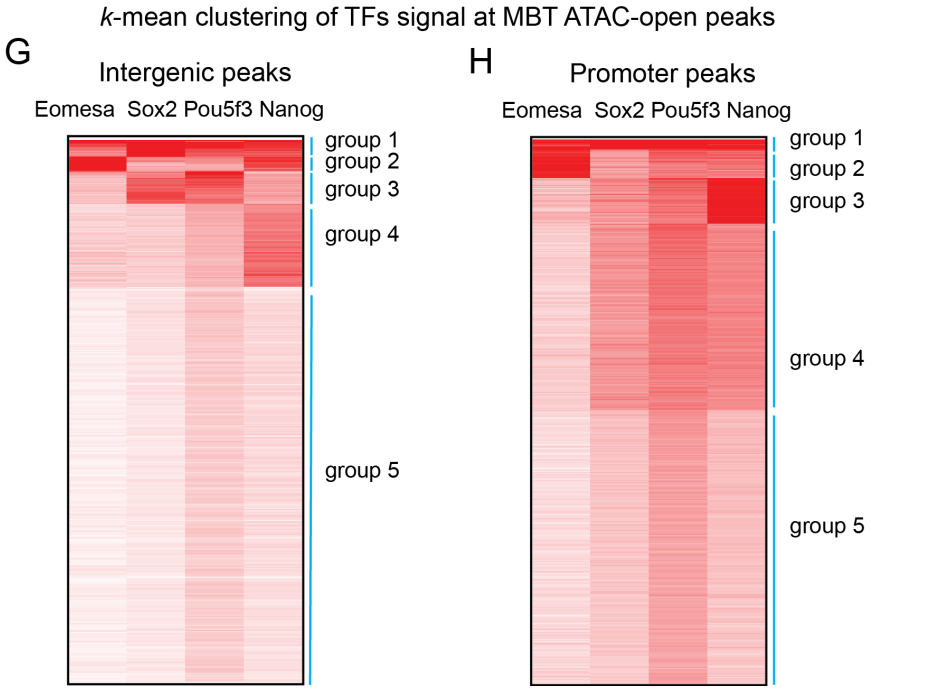

k-mean clustering of TFs signal at MBT ATAC-open peaks
G
Intergenic peaks
H
Promoter peaks
Eomesa Sox2 Pou5f3 Nanog
group 1
group 2
group 3
group 4
group 5

Figure 3.16: Eomesa, Sox2, Pou5f3 and Nanog bind both open and closed chromatin. (A) *k*-mean clustering of MBT ATAC-open signals at binding sites of Eomesa, Sox2, Pou5f3 and Nanog (from left to right). Potential open binding sites were marked by color bar besides each heatmap. (B) Pie chart of the percentage of potential open binding sites of each factors (Eomesa, Sox2, Pou5f3 and Nanog). (C) The mean profiles of MBT nucleosome signals around different Nanog binding sites ($\pm$2kb from the binding site center). (D) Same as in (C), for nucleosome signals at post-MBT.

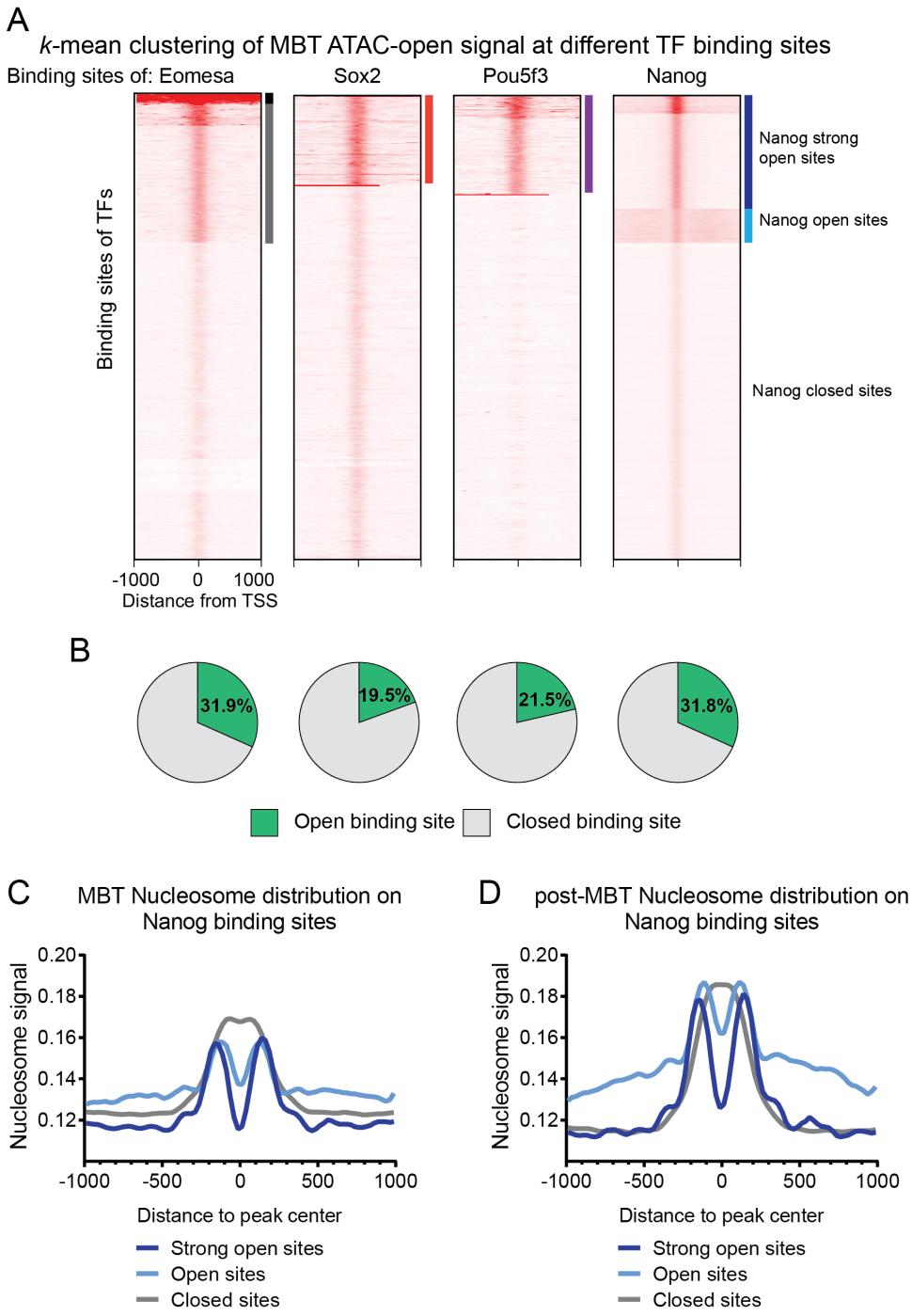
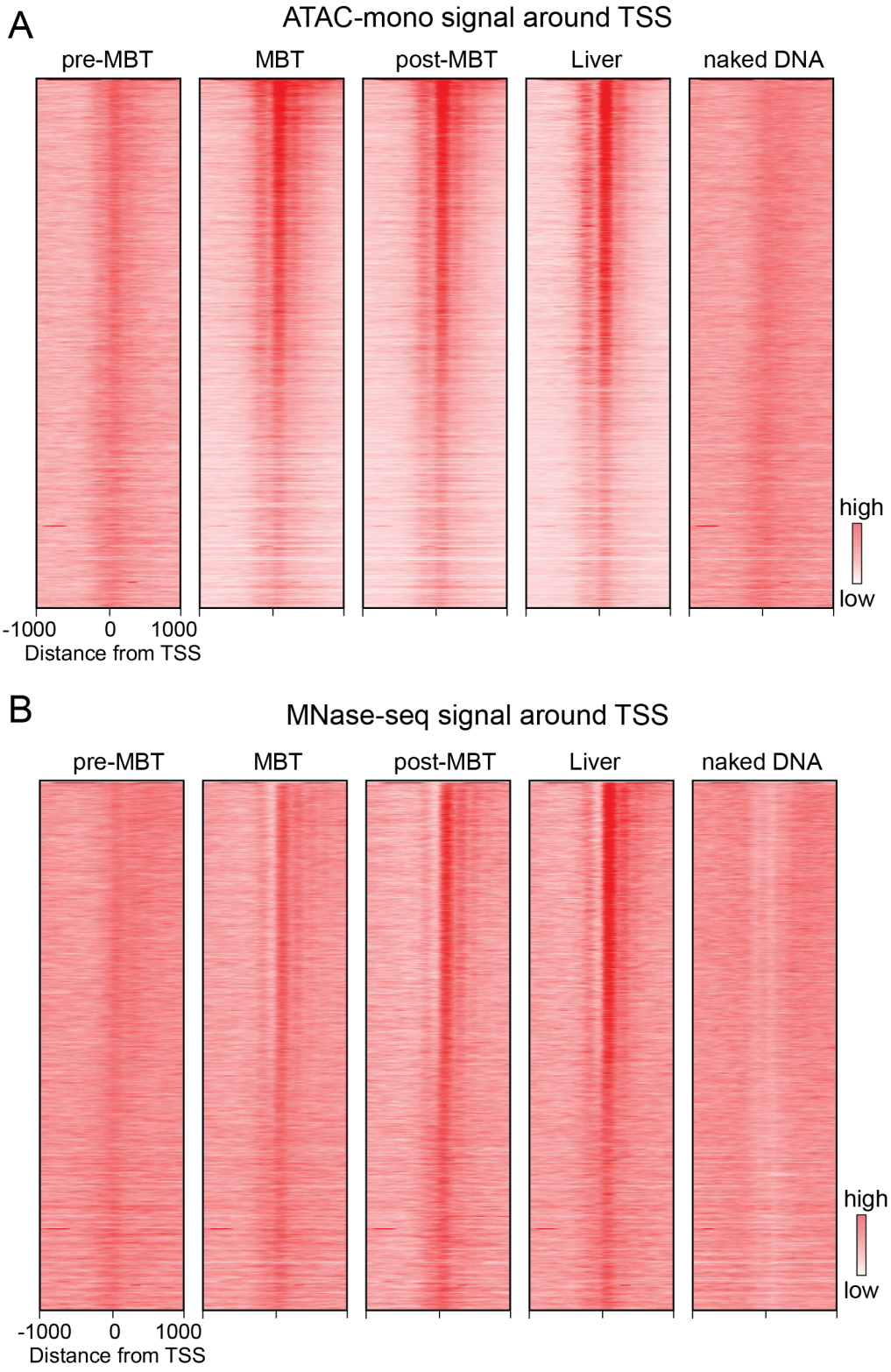

A
k-mean clustering of MBT ATAC-open signal at different TF binding sites
Binding sites of: Eomesa
Sox2
Pou5f3
Nanog
Nanog strong open sites
Nanog open sites
Nanog closed sites
Binding sites of TFs
-1000
0
1000
Distance from TSS
B
31.9%
19.5%
21.5%
31.8%
Open binding site
Closed binding site
C
MBT Nucleosome distribution on Nanog binding sites
Nucleosome signal
0.20
0.18
0.16
0.14
0.12
-1000
-500
0
500
1000
Distance to peak center
Strong open sites
Open sites
Closed sites
D
post-MBT Nucleosome distribution on Nanog binding sites
Nucleosome signal
0.20
0.18
0.16
0.14
0.12
-1000
-500
0
500
1000
Distance to peak center
Strong open sites
Open sites
Closed sites

Figure 3.17: ATAC-mono data can reflect nucleosome positioning around TSS. (A) Heatmaps of ATAC-mono signals of pre-MBT, MBT, post-MBT and liver around TSS ($\pm$1kb). TSS is ranked by gene expression at MBT (FPKM at 4hpf). (B) same as in (A), for MNase-seq data.


A
ATAC-mono signal around TSS
pre-MBT
MBT
post-MBT
Liver
naked DNA
high
low
-1000
0
1000
Distance from TSS
B
MNase-seq signal around TSS
pre-MBT
MBT
post-MBT
Liver
naked DNA
high
low
-1000
0
1000
Distance from TSS

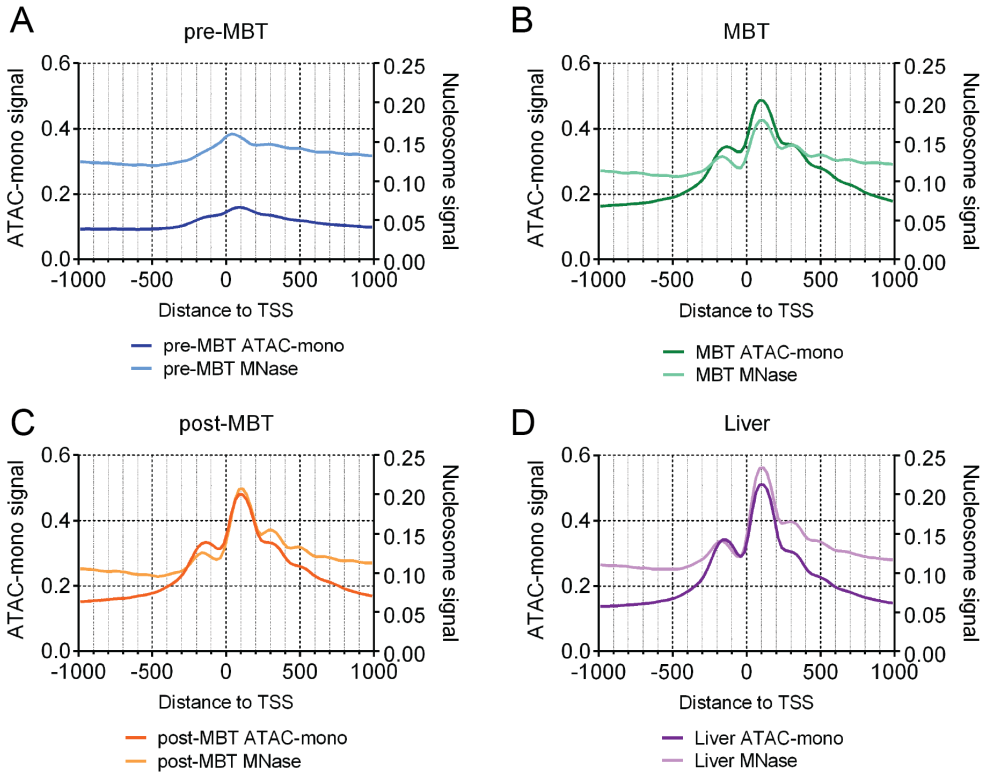


Figure 3.18: ATAC-mono cannot fully recapitulate nucleosome positioning compared to MNase-seq. The mean profiles of ATAC-mono and MNase-seq at each stage. pre-MBT (A), MBT (B), post-MBT (C), and Liver (D).

CHAPTER 4

NUCLEOSOME POSITIONING IN EARLY ZEBRAFISH EMBRYO

4.1 Abstract

Proper regulation of transcription initiation, elongation and co-transcriptional splicing involves dynamic changes in the presence and position of nucleosomes. Nucleosomes are the basic subunit of chromatin, which is composed of $\sim$147bp DNA and a histone octamer. Nucleosomes can inhibit transcription factor binding, while nucleosome-depleted DNA is highly accessible. ATAC-seq is useful for mapping the accessible chromatin regions, but is inadequate for determining the global nucleosome positioning. To better understand nucleosome occupancy and dynamics around the midblastula transition (MBT) and embryonic transcriptional onset, we employed MNase-seq during early stages of zebrafish embryogenesis. First, comparisons of nucleosome positioning from crosslinked and noncrosslinked samples at different stages generated similar nucleosome maps, with weak nucleosomes better captured under crosslinking conditions. At the genome-wide scale, nucleosome positioning from different stages was largely similar. However, at genes and promoters, nucleosome positioning from pre-MBT embryos is distinct from other stages. Interestingly, the formation of nucleosome arrays around TSS is independent of transcription. Instead, Pol II occupancy and DNA sequences

around the TSS appear to guide the establishment of a nucleosome array. Besides the TSS, certain exons also displayed high nucleosome occupancy – those that were correlated with polyA(5) sequences at their 5'ends followed by alternating A/T, or G/C dinucleotides inside the exons. These exons can be partitioned into two groups: 1) strong AT-GC alternation with infrequent polyA(5) sequences, or 2) weak AT-GC alternation with frequent polyA(5) sequences. Interestingly, we find these features and classifications of exons conserved in several other organisms, which may support current speculation that nucleosome positioning may help to coordinate RNA Pol II elongation and splicing.

4.2 Introduction

The nucleosome is the fundamental building block of a higher-order chromosome and is composed of 147bp DNA and a histone octamer (Kornberg, 1974). This DNA-histone interaction affects transcriptional activity by occlusion of DNA binding proteins, and leads to gene repression. Determining the nucleosome positioning in early embryos will help us to understand the chromatin landscape and how is it involved in gene regulation during embryogenesis. Additionally, the genome-wide profiling of nucleosomes also allows us to investigate to what extent DNA sequence instructs nucleosome positioning.

Nucleosome has been extensively studied for its role in gene regulation, which can be mediated through histone positioning, histone tail modifications and histone variants incorporation (Li et al., 2007). Nucleosome positioning describes where nucleosomes are located on the genomic DNA. When associated with nucleosomes, promoters and enhancers are inaccessible for most transcription factors

and co-regulators, usually leading to gene repression. Therefore, nucleosome positioning at regulatory elements are crucial for gene expression. The vast amount of post-transcriptional modifications on histone tails adds another level of complexity to transcription regulation. The modifications include methylation (mono-, di-, tri-methylation), acetylation, phosphorylation, ubiquitination, and newly discovered sumoylation and crotonylation (Tan et al., 2011). The diverse histone modifications function individually or cooperatively to convey active or repressive signals for gene expression. For instance, H3K4me1/3 and H3K27ac mark active promoters and enhancers, whereas H3K27me3 and H3K9me3 are indicators of gene repression and heterochromatin (Li et al., 2007). Furthermore, gene regulation is also influenced by the composition of histone octamer, where canonical histones (H2A, H2B, H3, and H4) can be replaced by histone variants. The H3 variants (H3.3, CenH3), and H2A variants (H2A.Z, H2A.X, macroH2A) are extensively studied and considered to be the most important. H3.3 marks active chromatin and is incorporated into mouse paternal genome immediately after fertilization (Santenard et al., 2010). H2A.Z is distributed at enhancers and upstream of transcription start site (TSS), and is thought to decrease the stability of nucleosome and thus facilitates chromatin accessibility (Jin et al., 2009).

Another essential question about nucleosome is how they are positioned throughout the genome. In principle, nucleosome positioning is determined by a combination of DNA-sequence, chromatin remodelers, transcription factors and RNA polymerases (Struhl & Segal, 2013). However, the relative importance of each of these determinants is under intense debate. One group argues DNA sequence plays a central role in determining the nucleosome position. Nucleosome affin-

ity to 147bp DNA varies based on the nucleotide compositions (Thåström et al., 1999). In general, DNA sequence with 10 bp periodic placement of A/T and G/C dinucleotides forms a nice curvature and makes stable contact with histones (Satchwell et al., 1986; Segal et al., 2006). In contrast, DNA composed of a strand of poly(dA:dT) is intrinsically stiff and is less feasible to wrap around the histone octamer, and in turn, often disfavors nucleosomes (Struhl & Segal, 2013). The sequence characteristics of nucleosome-favoring and -disfavoring sequence can be modeled mathematically and used to predict the likelihood of nucleosome formation at particular sequences (Segal et al., 2006). Surprisingly, the predicted nucleosome positioning is correlated with observed *in vivo* nucleosome positioning in yeast and *c.elegans* (Kaplan et al., 2009). Additionally, the *in vitro* reconstituted nucleosome maps resemble the endogenous nucleosome maps (Kaplan et al., 2009). These results strongly support the hypothesis that DNA plays a dominant role in positioning nucleosome. However, another group which also reconstituted nucleosome with yeast genomic DNA *in vitro* came to the opposite conclusion (Zhang et al., 2009). They only found moderate similarity between the *in vitro* and *in vivo* nucleosome map. Moreover, these two datasets differed significantly around TSS. An array of nucleosomes formed downstream of TSS in the *in vivo* data, which was absent in the *in vitro* data. Thus, they concluded DNA sequence contributes a minor role in nucleosome positioning. The discrepancy between the two studies probably stems from the preparation for *in vitro* reconstitution, the sequencing depth, and the computational analysis. Nevertheless, with the emergence of more nucleosome positioning data from other organisms, it is now generally accepted that DNA sequence contributes to nucleosome positioning, but

is not the dominant player (Valouev et al., 2008; Mavrich et al., 2008; Schones et al., 2008; Stein et al., 2010).

A more prevalent model for nucleosome positioning is: nucleosome depletion regions (NDR) are established at the poly (dA:dT) sequences, which are predominantly distributed at promoter regions. NDR is recognized by ATP-dependent chromatin remodelers (i.e. RSC, ISW2), which place -1 and +1 nucleosomes flanking the NDR. Transcription factors and the preinitiation complex further shape the unidirectional nucleosome array downstream of the +1 nucleosome. Finally, chromatin remodelers that assemble nucleosomes (i.e. ISWI and CHD1) are responsible for the nucleosome positioning at coding regions and other nongene regions (Struhl & Segal, 2013). In particular, the binding of preinitiation complex at TSS is essential for the establishment of the +1 nucleosome and downstream nucleosomes. On one hand, it can serve as a platform for the recruitment of chromatin remodelers. On the other hand, it has been proposed that the large complex can act as a barrier and permit the spacing of downstream nucleosomes, which is known as 'statistical positioning' (Kornberg & Stryer, 1988).

In this chapter, I described the genome-wide nucleosome occupancy at embryonic and differentiated stages of zebrafish. I found genome-wide nucleosome positioning is overall similar between early embryos and adult liver. However, the most variance is observed at gene and promoter regions between the pre-MBT sample and the rest of the samples. In addition, nucleosomes displayed a higher density at TSS and exons. The data suggest the formation of nucleosome array downstream of TSS is not dependent on transcription, but instead may be regulated by the Pol II occupancy and DNA sequence. I also revealed that exon

sequences displayed a high frequency of polyA(5) sequences at the 5'end of exons and A/T and G/C dinucleotides alternation inside exons, which might increase the exons-nucleosome affinity.

4.3 Results

4.3.1 Nucleosome positioning in early zebrafish embryos and adult liver

To understand how nucleosome positioning regulates gene activation in early zebrafish embryo, and whether nucleosome map in early embryo is different from that of differentiated tissue, I performed Micrococcal Nuclease (MNase) digestion on chromatin extracted from three embryonic stages (pre-MBT (4.5hpf), MBT (4hpf), post-MBT (5.3hpf), and adult liver (Figure 4.1). The digested mono-nucleosomes were then gel purified, and sent for paired-end Illumina sequencing (Table 4.1).

Although MNase digestion has been widely adopted for investigating nucleosome positioning, there is some variability and discrepancy across different experiments. To ascertain that our nucleosome maps truly reflect the endogenous nucleosome positioning, we have carefully controlled the experiments at three aspects. Firstly, over-digestion or under-digestion by MNase can impact the mapping of nucleosomes (Flores et al., 2014). Under-digestion generates a larger proportion of long fragments which will not be efficiently sequenced, thereby resulting in more regions uncovered. On the other hand, over-digestion leads to partial degradation of well-positioned nucleosomes, or even to dissociation of unstable nucleosomes. To avoid over-digestion and under-digestion, and ensure our digestions are comparable between stages, I have extensively optimized the procedures

to acquire a digestion pattern containing 70-80% of mono-nucleosomes, 10-20% of di-nucleosomes and some tri-nucleosomes (Figure 4.2). Secondly, the existing MNase digestion studies have used either crosslinked chromatin or noncrosslinked chromatin. Crosslinking covalently links the protein and DNA interaction and is less affected by experiment conditions (i.e. temperature, salt concentration, and detergents). However, crosslinking may introduce bias by linking unrelated protein-DNA interactions. Surprisingly, no report has directly compared nucleosome positioning maps generated from crosslinked (CL) and noncrosslinked (nCL) samples. Therefore, I conducted our MNase digestion at both crosslinked and noncrosslinked conditions, which allowed us to discover the endogenous nucleosome positioning and resolve the ambiguity. Thirdly, MNase preferentially cuts AT-rich sequence. To control for potential sequence bias introduced MNase, naked DNA was digested by MNase.

With these controls, our two replicates at each stage under two conditions (CL vs. nCL) are highly correlated (Pearson correlation around 0.8-0.94, Table 4.1). In line with the significant correlation, fragment size distributions between replicates are highly similar, except the ones for pre-MBT nCL (Figure 4.3). Nevertheless, the two pre-MBT nCL samples are highly correlated genome-wide (Figure 4.4A). Thus, I combined the two replicates at each stage at crosslinked and noncrosslinked conditions, respectively, for further analysis.

4.3.2 Comparison of crosslinked and noncrosslinked samples

Examining the fragment size distribution, I found crosslinked samples displayed a broader distribution compared to noncrosslinked samples, which could be be-

cause MNase is less efficient in cutting between crosslinked chromatin (Figure 4.3 A and B). The fragment size difference between CL and nCL samples was even more evident when comparing them side by side at each stage (Figure 4.3 C-F). To investigate whether the difference in size distribution represents the difference occupancy genome-wide, I collected the mean nucleosome signal every 100bp spanning the entire genome, and performed pairwise Pearson Correlation. Notably, CL and nCL samples were highly correlated (Figure 4.4 B-E). Also, I generated a predicted nucleosome map utilizing the computational model built by the Segal Lab with default parameters (c 0.1, t 1) (Kaplan et al., 2009). I argue that if CL and nCL are correlated, they should have a similar correlation to the predicted nucleosome positioning. Indeed, genome-wide correlation analysis indicated nucleosome maps from both crosslinked and noncrosslinked samples correlate with the predicted nucleosome map in a similar range (Table 4.2). In particular, digested fragments from chromatin, either CL or nCL, were significantly different from that of naked DNA (Table 4.2).

In addition to the genome-wide comparison, I also compared CL and nCL samples specifically at promoters. For post-MBT and liver, CL and nCL samples both formed nice nucleosome array around TSS, and are distinctly different from digested DNA (Figure 4.5C, D). In contrast, pre-MBT CL and nCL samples only exhibited moderate enrichments at TSS, yet no nucleosome array formed at the naked DNA sample (Figure 4.5A). Unexpected, MBT CL and nCL samples displayed different distribution around TSS (Figure 4.5B). MBT CL sample clearly captured the -1 nucleosome upstream of TSS, which was missing in nCL samples. Since -1 nucleosomes are less stable, this observation suggests CL samples may

better retain the less stable nucleosomes. In summary, CL and nCL were similar in general, but CL samples may capture weak nucleosomes better.

4.3.3 Dynamic nucleosome positioning at different stages

Nucleosome is the basic subunit of higher-order chromatin structure, which also protects the DNA from transposon insertion and inadvertent transcription. Since only around 2% of the genome codes for genes, nucleosome positioning should be largely unchanged globally in different cells or at development times, whereas gene related regions should display higher dynamics as a result of the cell-specific transcriptome. Indeed, when nucleosome positioning from two samples genome-wide were compared, the difference was mild under either CL or nCL condition (Figure 4.6A-F). Notably, at nCL condition, post-MBT and liver displayed higher variance compared to other samples, reflecting the larger difference of transcriptome between embryonic and differentiated tissue (Figure 4.6C, F). Hierarchical clustering is then applied to all correlations from any of two samples. nCL samples from embryonic stages were clustered as a group, and the rest of the samples were clustered into another group (Figure 4.6G). This suggests nucleosome signal from nCL embryonic samples are closely related. However, it has to be recognized the genome-wide difference is mild, as most the pairwise comparisons have Pearson Correlation score higher than 0.85.

To determine whether gene related areas have higher variation among samples at different stages, similar correlation heatmaps for gene regions and promoter regions were generated. Gene regions expand from upstream 5kb of gene start to downstream 5kb of gene end, while promoter regions represent 2kb around

TSS. Intriguingly, correlation at gene and promoter regions separated pre-MBT (CL and nCL) and MBT nCL samples from MBT CL, post-MBT and liver samples (Figure 4.7A, B). This indicates nucleosome positioning at gene related areas is different between early embryonic stage and later embryonic stage or differentiated samples.

4.3.4 Nucleosome array around transcription start site

Besides correlation, I also profiled the average nucleosome distribution around TSS. Nucleosome arrays (defined by +1 nucleosome after TSS, followed by downstream nucleosomes) were observed in MBT, post-MBT and liver samples, but not in the pre-MBT samples (Figure 4.5). Surprisingly, the predicted nucleosome signal also exhibited an array around TSS (Figure 4.8A). Because the predicted nucleosome signal is solely based on DNA sequence, I speculate DNA around TSS may have some sequence characteristics favoring nucleosome. Of note, the array in predicted nucleosome positioning shifted towards the upstream of TSS, and no -1 nucleosome was observed in predicted nucleosome positioning. These data suggest nucleosome formation around TSS is regulated by a combinatorial effect of DNA sequence and proteins, likely transcription machinery.

To determine whether the formation of nucleosome array indeed relates to gene transcription, genes were ranked based on the expression level (FPKM value from RNA-seq) at MBT (4hpf). Interestingly, highly transcribed genes displayed a better nucleosome array at MBT, which was more prominent at post-MBT (Figure 4.8B). The same trend was observed when genes were grouped into four quantiles based on their expression at MBT (Figure 4.9). For MBT, post-MBT and liver, the 1st quan-

tile genes (top 25% transcribed) have distinct nucleosome depleted regions (NDR), strong -1 nucleosomes, and nucleosome array downstream TSS. However, these features are missing in the 4th quantile genes (bottom 25% transcribed). Curiously, nucleosomes in liver also formed better nucleosome array at genes that are highly transcribed at MBT (Figure 4.8B, Figure 4.9D). This can be interpreted in two ways. First, the genes that are highly transcribed at MBT are also expressed at a high level in liver. Thus nucleosome array formation is dependent on transcription. Second, the highly transcribed genes at MBT have nucleosome favoring sequences around the TSS, thereby nucleosome array consistently formed regardless the cell types.

RNA-seq at MBT detects mRNA abundance regardless whether it is maternally deposited or newly transcribed, and therefore is not perfect to use as an indicator for transcription level. To faithfully reflect gene transcription status, I resorted to RNA Pol II occupancy determined by ChIP-seq. Genes were grouped into Pol II bound (indicating poised and transcribed genes) and non-Pol II bound (implicated nontranscribed genes). One hundred sixty genes are bound by Pol II at pre-MBT, which increased substantially to 5218 genes at MBT and 6746 genes at post-MBT. Interestingly, the genes bound by Pol II at pre-MBT have better nucleosome array compared to the unbound genes, which is consistently observed at MBT and post-MBT (Figure 4.10). In contrast, Pol II binding failed to differentiate the predicted nucleosome distribution, supporting the hypothesis that transcription (indicated by Pol II binding) is required for the nucleosome positioning around promoter.

4.3.5 Nucleosome array formation at promoter is not dependent on transcription

To prove transcription regulates nucleosome array formation, I inhibited transcription initiation by injecting a-amanitin in 1-cell embryo or inhibited transcription elongation by soaking embryo in embryo water containing flavopiridol. The a-amanitin and flavopiridol treatments were highly efficient, as nearly all treated (both a-amaintin and flavopiridol) embryos were arrested at MBT. Then, MNase-seq were employed on the chromatins of drug-treated embryos and their controls. Unexpectedly, nucleosome array still formed on Pol II-bound genes even when transcription initiation or elongation was inhibited (Figure 4.11A-D). These data suggest the transcription activities are not the determinant of nucleosome array formation; rather it is Pol II binding that instructs the nucleosome positioning. It has long been speculated that Pol II binding serves as a barrier allowing 'statistical positioning'. This theory suggests that once a boundary is established, other nucleosomes will naturally form around that boundary following thermodynamic equilibrium. Here, our data support this model.

In yeast, DNA around TSS often harbors polyA sequences which create nucleosome depletion regions that are crucial for +1 and -1 nucleosome placement. Therefore, it is possible that Pol II-bound genes have a higher frequency of polyA track than that of unbound genes, which allows better nucleosome formation around TSS. The possibility might also be applied to highly transcribed genes. To test, I profiled the frequency of AAAAA sequence (A5mer) around TSS. Surprisingly, there is no difference in A5mer sequence between Pol II bound and unbound genes, or between highly transcribed and poorly transcribed genes (Figure 4.12A).

I then turn to nucleosome favoring sequence, which often displayed $\sim$10bp

periodicity of AT and GC dinucleotides. The AT-GC alternation renders the DNA form a nice curvature, which can make stable contact with histones. Similar to A5mer sequence, the frequency of AT and GC dinucleotide around TSS were calculated. Interestingly, genes that are highly transcribed or bound by Pol II at MBT have a higher GC content at the promoter and exhibit a better AT and GC dinucleotide alternation compared to genes poorly transcribed or not bound by Pol II at MBT (Figure 4.12 B-E). Notably, a AT dinucleotide peak is located around -30bp upstream of TSS, which is coincident with TATA box. This suggests TATA box in zebrafish gene could serve as a poly(dA:dT) sequence and play a role in nucleosome positioning. Taken together, genes that are highly transcribed or bound by Pol II at MBT have moderately higher intrinsic nucleosome affinity around the TSS, and Pol II binding further enhances the formation of downstream nucleosomes.

4.3.6 Nucleosome enrichment over exons

Nucleosome has been reported enriched on exons (Tilgner et al., 2009). To test whether this is true in zebrafish, I profiled nucleosome signal over exons grouped by their lengths. The majority of the zebrafish exons are about 100-300bp long, which is close to a nucleosome size, $\sim$147bp (Figure 4.13A). The nucleosomal DNA wraps around histone and contacts histones every 11bp as the minor groove faces the protein. As shown in Figure 4.13B, I defined nucleosome-size exons with a length of 125bp to 169bp. To match the 44bp length difference in this exon group, I set the lower cutoff for small exons as 81bp and the upper cutoff for large exons as 213bp. Then, the average nucleosome signals around the center of

exons of all three groups were profiled. Consistent with the observations in other organisms, nucleosomes of all examined stages were enriched at exons, as well as the predicted nucleosomes. The high enrichment of predicted nucleosome over exons suggests exons might have nucleosome favoring sequences. Indeed, the 5'end, but not the 3'end, of all 3 groups of exons have a high frequency of polyA sequence (Figure 4.14). Interestingly, nice alternation of AT, GC dinucleotides is observed over exons, which is absent from the neighboring sequence (Figure 4.15). In sum, the data suggest polyA tracks present at the 5'end of exons may establish a boundary for nucleosome positioning, and inside the exons, AT and GC alternation further promote the stable interactions between DNA and histones. Therefore, together they contribute to stabilize nucleosome over exons.

4.3.7 Sequence characteristics of exons

To determine whether some exons have a better AT and GC alternation than others, *k*-mean clustering was performed on each exon group. Interestingly, exons were clustered into subgroups based on AT and GC alternation for nucleosome-size exons (125-169bp) and large exons (169-213bp), but not for small exons (81-125bp) (Figure 4.16). Then, I profiled the AT, GC dinucleotides and A5mer signals for each subgroup of nucleosome-size exons. Strikingly, the top three subgroups showed a very nice AT-GC alternation. In contrast, the bottom three subgroups had poor AT-GC alternation but had slightly higher frequency of A5mer present on the 5'end (Figure 4.17). Similar results were observed for subgroups of large exons (data not shown).

To understand whether the nucleosome positioning sequence on exons is con-

served, A5mer, AT and GC distributions were analyzed on exons of human, mouse, fly, medaka, fugu, and pufferfish. Intriguingly, this phenomenon is conserved in other species: small exons have poor AT-GC alternation, nucleosome-size exons and large exons have better AT-GC alternation, which can be divided into subgroups using *k*-mean clustering. As examples, human and mouse subgroups of nucleosome-size exons (125-169bp) were shown in Figure 4.18. Both human and mouse subgroup 1 have very nice AT-GC alternation, and they have a weak A5mer sequence at the 5'end of exons. In contrast, subgroup 6 exons of human and mouse have poor AT, GC alternation, but have a strong A5mer sequence at the 5'end of exons. This suggests different exons employ different strategies to position nucleosome. Some exons, such as the one in subgroup 1, position nucleosome predominantly through the AT-GC alternation sequence. Other exons, such as the ones in subgroup 6, position nucleosome preferentially through the polyA sequence.

4.4 Discussion

The nucleosome is the basic subunit of chromatin, which is composed of $\sim$147bp DNA and eight histone proteins (two copies of H2A, H2B, H3, and H4). Each nucleosome is separated by linker DNA (20-80bp), which is occupied by the fifth histone, H1, which facilitate the condensation of nucleosomes into higher-order chromatin structure (Kornberg, 1977; Luger et al., 1997). Besides its role in genome compaction, the role of nucleosome in gene regulation has also been intensively studied, which involves histone variants, histone modifications and nucleosome positioning. In this chapter, I have described the profiles of nucleosome positioning

in early embryos and adult liver cells, and provided insights on how nucleosomes are positioned at promoters and exons.

At promoters, nucleosome arrays are observed in MBT, post-MBT, liver but not pre-MBT samples. The nucleosome arrangement was also found in predicted nucleosome but was distinct from the ones in observed nucleosome maps. Predicated nucleosomes had no -1 nucleosome and NDR, and the center of +1 nucleosome was shifted upwards and resides right at TSS. In observed nucleosome profiles, the center of +1 nucleosome was downstream of TSS. This suggests nucleosome array does not solely rely on DNA sequence, and other factors are involved. However, nucleosome array around TSS was not dependent on transcription, as no difference is observed between untreated embryos and embryos treated with a-amanitin and flavopiridol (transcription inhibitors). Interestingly, genes bound by Pol II formed a stronger nucleosome array compared to genes that are not bound by Pol II. This was true in pre-MBT, MBT and post-MBT embryos. However, predicted nucleosome showed no difference in the two gene sets, indicating their promoter DNA sequence were largely the same. Therefore, I speculate Pol II occupancy may result in different nucleosome distributions at promoters. This speculation can be supported by the ”statistical positioning” model which hypothesizes that boundary (specific DNA sequence and/or protein binding) leads to an array of evenly spaced nucleosome (Kornberg, 1981; Kornberg & Stryer, 1988).

The next question is how Pol II selectively binds those genes,; could the underlying DNA sequence play an instructive role? To test, I focused on two types of sequences, A5mer and AT-GC alternation. A 5nt long DNA has 1024 potential sequence combinations. Among these, A5mer sequence has the lowest nucle-

osomes density in yeast. The related poly(dA:dT) sequences, which are highly enriched at the promoters of yeast genes, are proposed to create a boundary for nucleosome array formation. Surprisingly, I do not observe a high frequency of A5mer around TSS. It is possible that by counting the Amer using a fixed windows (5bp) starting always at TSS, I might miss the A5mers that expand across two windows. I can test this by generating different frequency tracks which start at TSS, TSS+1bp, TSS+2bp, TSS+3bp, and TSS+4bp. In this way, all potential A5mers can be captured by combining all these five tracks. An alternative explanation is that other poly(dA:dT) sequences rather than A5mer are used in fish to create nucleosome depleted regions. Besides A5mer, the distribution of A/T and G/C dinucleotides was profiled. A peak of A/T dinucleotides around -30bp of TSS is observed for all genes, which is coincident with where TATA boxes reside, and may serve as a boundary. Moreover, I observed AT-GC alternation is better around TSS of Pol II bound genes comparing to genes that are not bound by Pol II. Based on the known higher affinity of AT-GC alternating sequences, I speculate the AT-GC alternation sequence might play a role in nucleosome formation. Together with Pol II occupancy, they promote the nucleosome array establishment.

Another interesting finding is that exons have sequence characteristics that enhance their affinity for nucleosomes. Specifically, high frequency of A5mer sequence is observed right at the 5'end but not 3'end boundary of exons, while inside exons, higher GC content, and AT-GC alternation are evident. Based on the sequence features, exons can be further separated into two groups. One contains better AT-GC alternation inside exons, while the other displays higher frequency of A5mer at the 5'end of exons. The grouping of exons can be applied

to other vertebrates, including human, mouse, medaka, pufferfish and fugu, but is not so applicable to fly. This suggests a conserved strategy in vertebrates for positioning nucleosome on exons. Furthermore, among all the organism that have been examined, the difference of GC content between exons and adjacent area is greater in fish (zebrafish, medaka, pufferfish, fugu) compared to mammals (human and mouse).

Exons may utilize a high level of nucleosome to prevent internal transcription and mutations introduced by transposons insertions. Alternately, high enrichment of nucleosome over exons may facilitate splicing (Tilgner et al., 2009). Pol II slows down when confronting exons with a high density of nucleosomes (Lowary & Widom, 1998; Bondarenko et al., 2006). It is also known that splicing can take place co-transcriptionally. Thus, the temporarily decelerated Pol II could gain time for the spliceosome to conduct splicing. Surprisingly, splicing can also affect nucleosome density. Over-expression of U1 snRNA or strengthening the 5'splice site increases the nucleosome density over exons while inhibiting splicing with meayamycin also alters nucleosome density (Keren-Shaul et al., 2013). Taken together, the enrichment of nucleosome on exons is driven by specific sequence characteristics and may protect exon from mutation and undesired transcription, as well as contribute to splicing regulation.

In conclusion, my work in profiling chromatin landscape by MNase-seq provides valuable information on nucleosome positioning at different stages, and also is a complementary to ATAC-seq data that focused mainly on open chromatin. The nucleosome positioning data allow us to discover interesting sequence characteristics at promoters and exons.

4.5 Methods

4.5.1 Zebrafish and sample collection

Tubingen zebrafish lines were maintained and raised under standard conditions. Wildtype embryos were collected after 10 min of mating to ensure the synchrony of embryo development, raised in embryo water at 28°C and staged as previously described (Kimmel et al., 1995).

4.5.2 Nuclei extraction from crosslinked embryonic samples

Embryos at pre-MBT (2.5hpf), MBT (4hpf) and post-MBT (5.3hpf) were carefully staged, dechorinated by Pronase (Sigma-Aldrich 11459643001), and fixed in 1% formaldehyde for 10 min at Room Temperature (RT). Formaldehyde was quenched by adding glycine to a final concentration of 0.125M, followed by centrifugation at 500g 5 min 4°C. The embryo pellet was resuspended in ice-cold PBS, and centrifuged again at 500g 5 min 4°C. The washing was repeated twice, and the embryos were snap frozen in liquid nitrogen, until enough embryos were collected. For chromatin extraction, embryo pellets were resuspended in cell lysis buffer (10mM Tris-HCl [pH 8.1], 10mM NaCl, 0.5% NP40) and lysed for 10 min, rotating at 4°C. Nuclei were collected by centrifugation at 1300g 5 min, 4°C, and washed twice by rotating at RT 5 min in nuclei wash buffer (50mM Tris-HCl [pH 8.0], 100mM NaCl, 10mM EDTA).

4.5.3 Nuclei extraction from noncrosslinked embryonic samples

Embryos at pre-MBT (2.5hpf), MBT (4hpf) and post-MBT (5.3hpf) were carefully staged, and rinse with ice-cold PBS. Cell lysis buffer (10mM Tris-HCl [pH 8.1],

10mM NaCl, 0.5% NP40) is add, and lysed for 10 min on ice. During the 10 min lysis, a 20 Gauge syringe was used three times to break the chorions and embryonic cells, followed by table top spin for 10 seconds. The supernatant was then centrifuged at 1300g 5 min 4°C and the nuclei pellet was ready.

4.5.4 Nuclei extraction from crosslinked and noncrosslinked liver cells

Adult fish were fasted for a week prior to dissection. Euthanasia of zebrafish with ice water was conducted following the IACUC procedures. Dissection of the adult zebrafish liver was performed under brightfield imaging on a dissection microscope. Dissected livers were immediately transferred to Eppendorf tubes on ice and rinsed with PBS.

For crosslinking liver cell, liver tissue was homogenized and fixed in 1% formaldehyde for 10 min at RT. Formaldehyde was quenched by adding glycine to a final concentration of 0.125 M, followed by centrifugation at 500g 5 min 4°C. The liver cell pellet was resuspended in ice-cold PBS, and centrifuged again at 500g 5 min 4°C. The washing was repeated twice. Liver cell pellet was resuspended in cell lysis buffer (10mM Tris-HCl [pH 8.1], 10mM NaCl, 0.5% NP40) and lysed for 10 min, rotating at 4°C. During the 10 min lysis, a 20 Gauge syringe was used 10 times, and the cell suspension was frequently pipetted. Nuclei were collected by centrifugation at 1300g 5 min, 4°C.

Noncrosslinking liver cells were treated essentially the same as the crosslinked sample, but were not crosslinked.

4.5.5 Drug treatment

For inhibition of transcription initiation, 1nl of a-amanitin (Millipore 129741-1MG) with concentration of 1ug/ml was injected into 1-cell embryos. Embryos without injection were used as control. For inhibition of transcription elongation, embryos were dechorinated immediately after fertilization, then flavopridol (Sigma F3055-1mg) was added to a final concentration of 1.5uM. Embryo incubated with 0.1% DMSO were used as control. The drug treated embryos were arrested around 4hpf. Control and drug treated embryos were then collected and the procedures in 4.5.2 for nuclei extraction were followed.

4.5.6 MNase digestion

Nuclei were resuspended in MNase digestion buffer (50mM NaCl, 50mM Tris [pH 7.4], 5mM MgCl2, 1mM CaCl2, 0.1% triton), followed by centrifugation at 1300g 5 min, 4°C. The nuclei were then resuspended in 500ul MNase digestion buffer, and equilibrated at RT for 5 min. The nuclei were then incubated with MNase (Affymetrix, 70196Y) at 37°C for 15 min. The MNase digestion was quenched by adding 56ul 10X stop buffer (50mM Tris [pH 7.5], 500mM NaCl, 100mM EDTA, 20mM EGTA, 10% Triton X-100). The digested chromatins then were RNase treated (Ambion, AM2286) for 2hr at thermomixer (37°C 400rpm 2hr), followed by 2hr proteinase K treatment at 55°C, reverse crosslinked overnight at 65°C (skipped for noncrosslinked samples). The DNA was purified with Phenol:chloroform pH8 (Invitrogen, No 15593-031), and was ethanol precipitated. DNA corresponding to mono-nucleosome size was then gel purified.

4.5.7 Library preparation and sequencing

Libraries of ChIP-seq samples were prepared using the NEBNext ChIP-Seq Library Prep Master Mix (NEB, E6240L). Eight ng of DNA was used and followed the manufacturer's procedure. Samples were then amplified by NEBNext High-Fidelity 2X PCR Master Mix (NEB, M0541L) and sequenced paired-end 101nt on an Illumina HiSeq 2000. For mono-nucleosomes from drug treated and control samples, 50nt single-end sequencing were performed.

4.5.8 Data analysis

Paired-end fastq files were uniquely aligned to zv10 zebrafish genome assembly using Novoalign (Novocraft, Inc) with the following parameters: `-r None -Q 13 -k -o SAM -a CTGTCTCTTATACACATCT`, and converted to BAM files using Samtools. BAM files from replicates were merged and bigwig files were generated using the program bam2wig (`http://search.cpan.org/dist/Bio-ToolBox/`) with the following parameters: `--position span --pe --rpm --bw`.

4.6 Contribution and acknowledgement

This work in combination with the work in Chapter 3 are under preparation for publication. I have performed all the experiments and bioinformatic analysis. I want to thank Tim Parnell and Tim Mosbruger for bioinformatics assistance, and Brian Dalley for sequencing expertise.

4.7 References

Bondarenko, V.A., Steele, L.M., Újvári, A., Gaykalova, D.A., Kulaeva, O.I., Polikanov, Y.S., Luse, D.S., Studitsky, V.M. (2006). Nucleosomes can form a polar barrier to transcript elongation by rna polymerase ii. Molecular Cell *24*, 469–479.

Flores, O., Deniz, Ö., Soler-López, M., Orozco, M. (2014). Fuzziness and noise in nucleosomal architecture. Nucleic Acids Research *42*, 4934–4946.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., Felsenfeld, G. (2009). H3. 3/h2a. z double variant–containing nucleosomes mark'nucleosome-free regions' of active promoters and other regulatory regions. Nature genetics *41*, 941–945.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al. (2009). The dna-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Keren-Shaul, H., Lev-Maor, G., Ast, G. (2013). Pre-mrna splicing is a determinant of nucleosome organization. PLoS One *8*, e53506.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Developmental Dynamics *203*, 253–310.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and dna. Science *184*, 868–871.

Kornberg, R.D. (1977). Structure of chromatin. Annual Review of Biochemistry *46*, 931–954.

Kornberg, R. (1981). The location of nucleosomes in chromatin: specific or statistical. Nature *292*, 579–580.

Kornberg, R.D., Stryer, L. (1988). Statistical distributions of nucleosomes: non-random locations by a stochastic mechanism. Nucleic Acids Research *16*, 6677–6690.

Li, B., Carey, M., Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Lowary, P., Widom, J. (1998). New dna sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. Journal of Molecular Biology *276*, 19–42.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 å resolution. Nature *389*, 251–260.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., et al. (2008). Nucleosome organization in the drosophila genome. Nature *453*, 358–362.

Santenard, A., Ziegler-Birling, C., Koch, M., Tora, L., Bannister, A.J., Torres-Padilla, M.E. (2010). Heterochromatin formation in the mouse embryo requires critical residues of the histone variant h3. 3. Nature Cell Biology *12*, 853–862.

Satchwell, S.C., Drew, H.R., Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core dna. Journal of Molecular Biology *191*, 659–675.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.P.Z., Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Stein, A., Takasuka, T.E., Collings, C.K. (2010). Are nucleosome positions in vivo primarily determined by histone–dna sequence preferences? Nucleic Acids Research *38*, 709–719.

Struhl, K., Segal, E. (2013). Determinants of nucleosome positioning. Nature Structural & Molecular Biology *20*, 267–273.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell *146*, 1016–1028.

Thåström, A., Lowary, P., Widlund, H., Cao, H., Kubista, M., Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning dna sequences. Journal of Molecular Biology *288*, 213–229.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nature Structural & Molecular Biology *16*, 996–1001.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. (2008). A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. Genome Research *18*, 1051–1063.

Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., Struhl, K. (2009). Intrinsic histone-dna interactions are not the major determinant of nucleosome positions in vivo. Nature Structural & Molecular Biology *16*, 847–852.

Table 4.1: MNase sequencing summary

	Pairs	Pairs aligned	% Aligned	Correlation
pre-MBT_CL_rep1	37,490,039	33,735,777	89.99%	0.9369
pre-MBT_CL_rep2	32,977,143	29,604,460	89.77%	
pre-MBT_nCL_rep1	35,331,450	32,105,355	90.87%	0.9460
pre-MBT_nCL_rep2	36,130,316	31,084,058	86.03%	
MBT_CL_rep1	30,258,189	27,556,610	91.07%	0.9355
MBT_CL_rep2	28,785,987	26,107,065	90.69%	
MBT_nCL_rep1	45,215,194	31,646,875	69.99%	0.8027
MBT_nCL_rep2	24,379,781	22,281,361	91.39%	
post-MBT_CL_rep1	37,363,896	33,990,207	90.97%	0.93004
post-MBT_CL_rep2	34,434,974	31,407,138	91.21%	
post-MBT_nCL_rep1	33,673,614	30,779,492	91.41%	0.9754
post-MBT_nCL_rep2	31,924,539	29,087,609	91.11%	
Liver_CL_rep1	27,887,876	25,516,455	91.50%	0.9392
Liver_CL_rep2	28,589,271	26,095,727	91.28%	
Liver_nCL_rep1	25,380,483	23,199,032	91.41%	0.8683
Liver_nCL_rep2	27,573,002	25,127,951	91.13%	
Naked_DNA(single-end)	84,740,186	62,303,818	73.50%	

Table 4.2: Correlation between observed and predicted nucleosome map

	Predicted nucleosome map	Digested Naked DNA
pre-MBT CL	0.3096891	0.2783015
pre-MBT nCL	0.4936699	0.2035423
MBT CL	0.3413961	0.2097353
MBT nCL	0.4957017	0.2931172
post-MBT CL	0.5262779	0.2297756
post-MBT nCL	0.4719718	0.1989889
Liver CL	0.3713786	0.2544334
Liver nCL	0.3541750	0.2181286

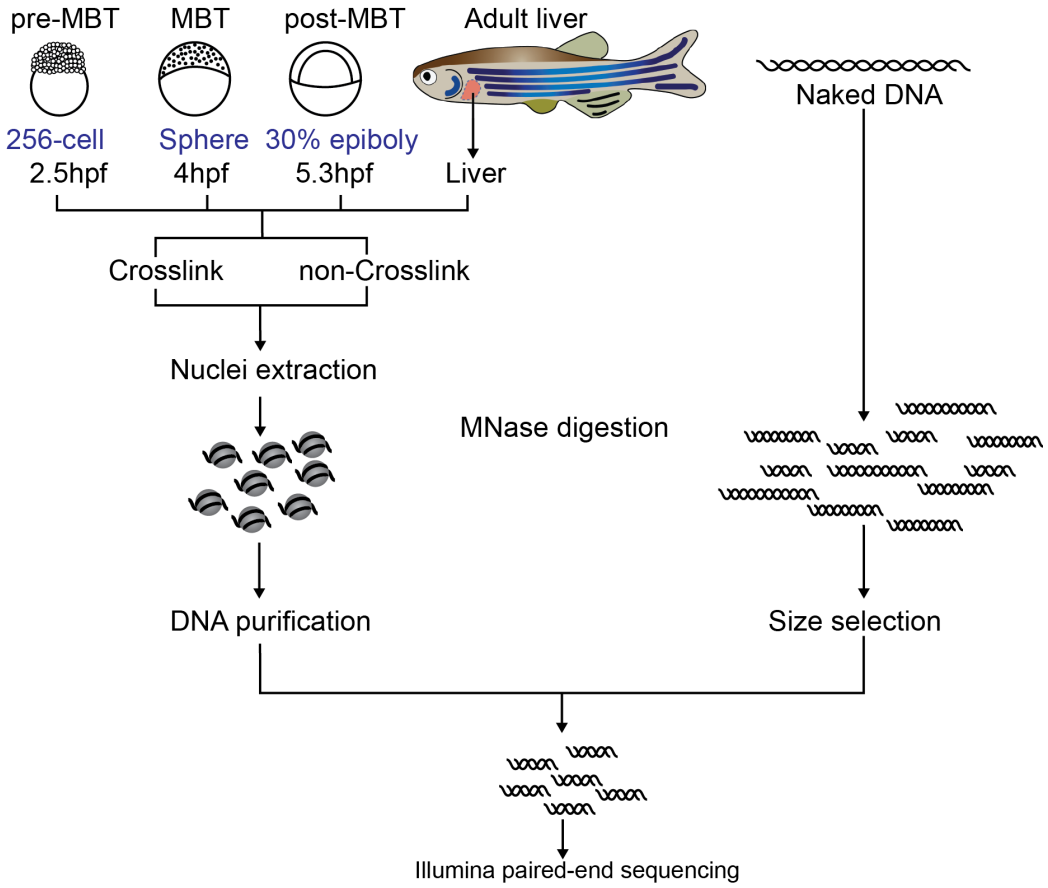
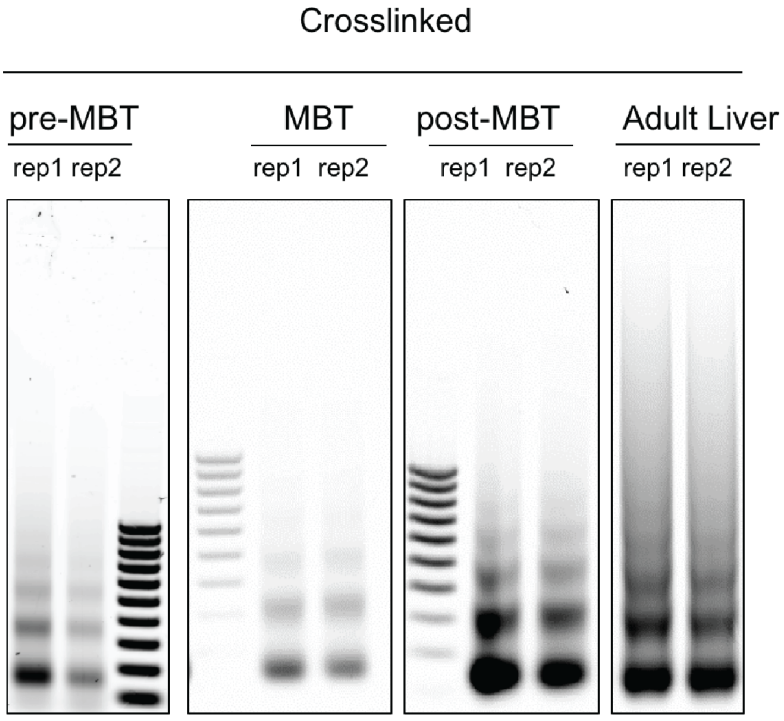


Figure 4.1: Schematic of MNase-seq data in zebrafish embryos at pre-MBT, MBT and post-MBT stages, and adult liver. At each stage, crosslinked and noncrosslinked chromatin are extracted in parallel, and are subject to MNase-digestion. Mono-nucleosomes are gel extracted, and the associated DNA is purified and sequenced. Naked_DNA is also MNase-digested and sequenced and serves as a control.

Figure 4.2: Agarose gel of MNase digestion. (A) MNase digestion of crosslinked samples. Two replicates from each stage are digested and separated by electrophoresis on aragarose gels. (B) Similar to (A) but for noncrosslinked samples.

A


Crosslinked
pre-MBT
rep1 rep2
MBT
rep1 rep2
post-MBT
rep1 rep2
Adult Liver
rep1 rep2

B

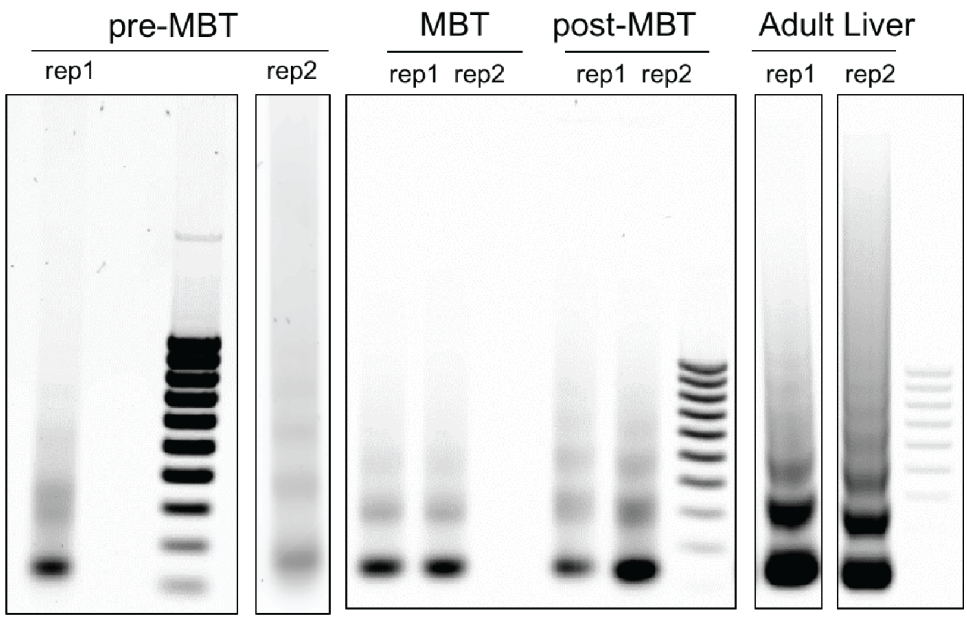
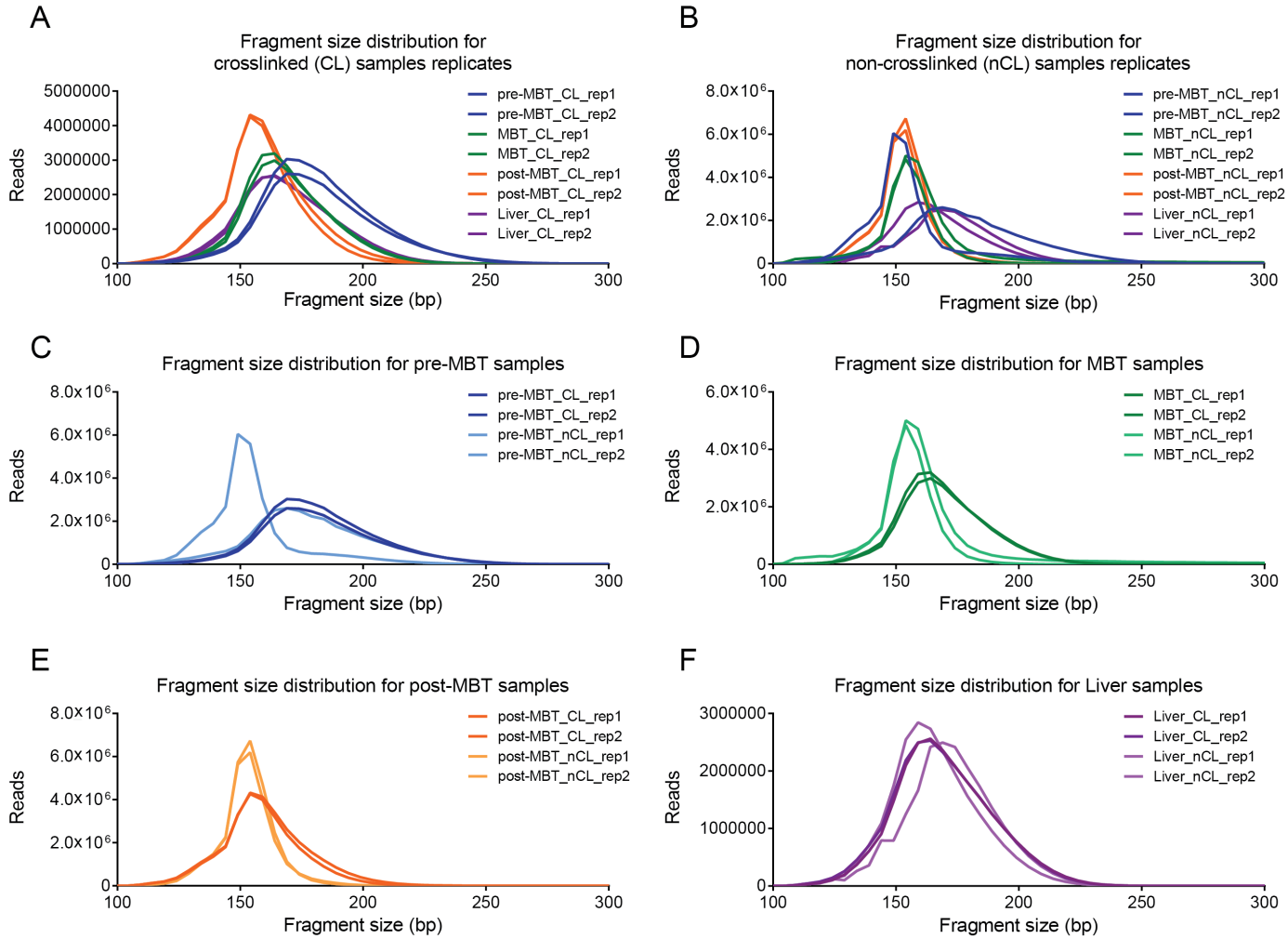

Non-crosslinked
pre-MBT
rep1
rep2
MBT
rep1 rep2
post-MBT
rep1 rep2
Adult Liver
rep1
rep2

Figure 4.3: Fragment size distribution. (A-F) Fragment size distribution for all crosslinked (CL) samples (A), all noncrosslinked (nCL) samples (B), CL and nCL replicates at pre-MBT (C), CL and nCL replicates at MBT (D), CL and nCL replicates at post-MBT (E), and CL and nCL replicates at Liver (F). The sizes of sequenced fragments are acquired from pair-end sequencing.


A
Fragment size distribution for crosslinked (CL) samples replicates
Reads
Fragment size (bp)
pre-MBT_CL_rep1
pre-MBT_CL_rep2
MBT_CL_rep1
MBT_CL_rep2
post-MBT_CL_rep1
post-MBT_CL_rep2
Liver_CL_rep1
Liver_CL_rep2
B
Fragment size distribution for non-crosslinked (nCL) samples replicates
pre-MBT_nCL_rep1
pre-MBT_nCL_rep2
MBT_nCL_rep1
MBT_nCL_rep2
post-MBT_nCL_rep1
post-MBT_nCL_rep2
Liver_nCL_rep1
Liver_nCL_rep2
C
Fragment size distribution for pre-MBT samples
D
Fragment size distribution for MBT samples
E
Fragment size distribution for post-MBT samples
F
Fragment size distribution for Liver samples

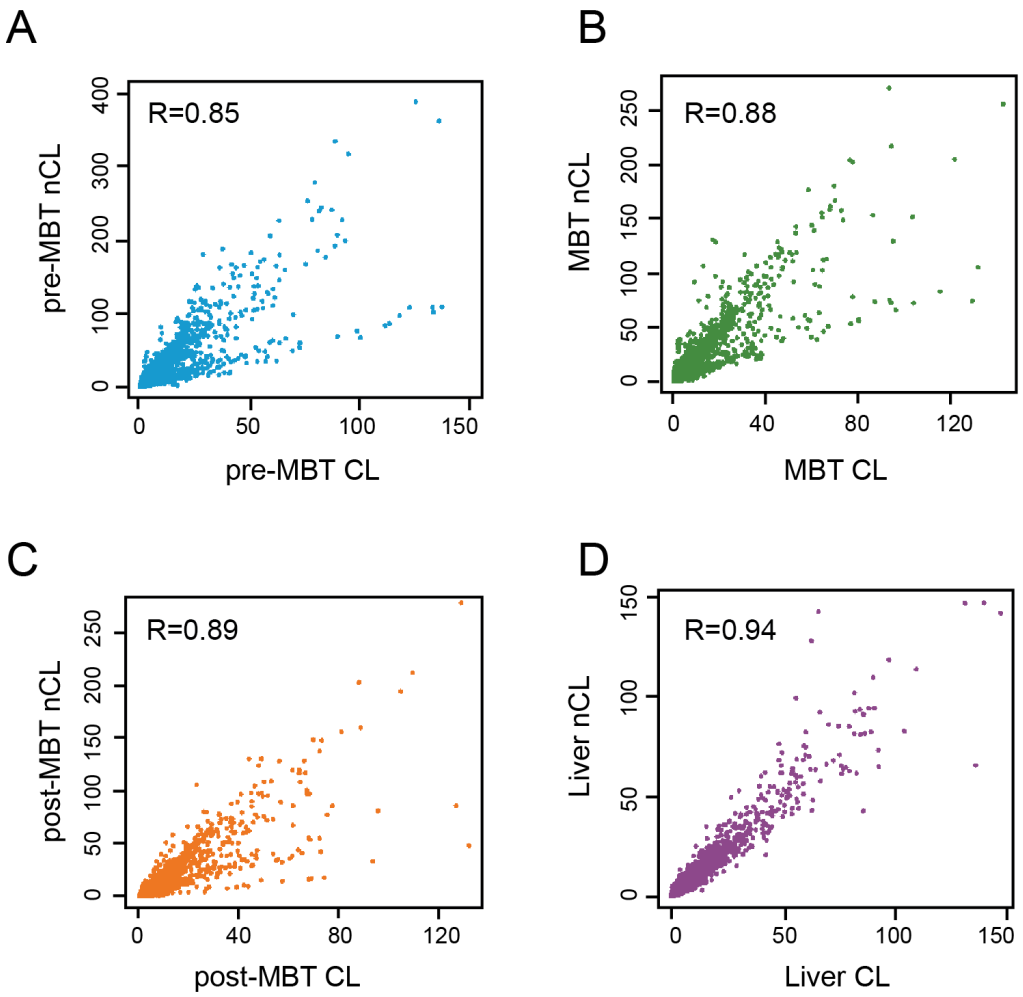


Figure 4.4: Genome-wide correlation between crosslinked and noncrosslinked samples at each stage. (A-D) Pairwise comparisons of normalized nucleosome signal per 100bp in CL and nCL samples at pre-MBT (A), MBT (B), post-MBT (C) and Liver (D). R value represents Pearson Correlation score.

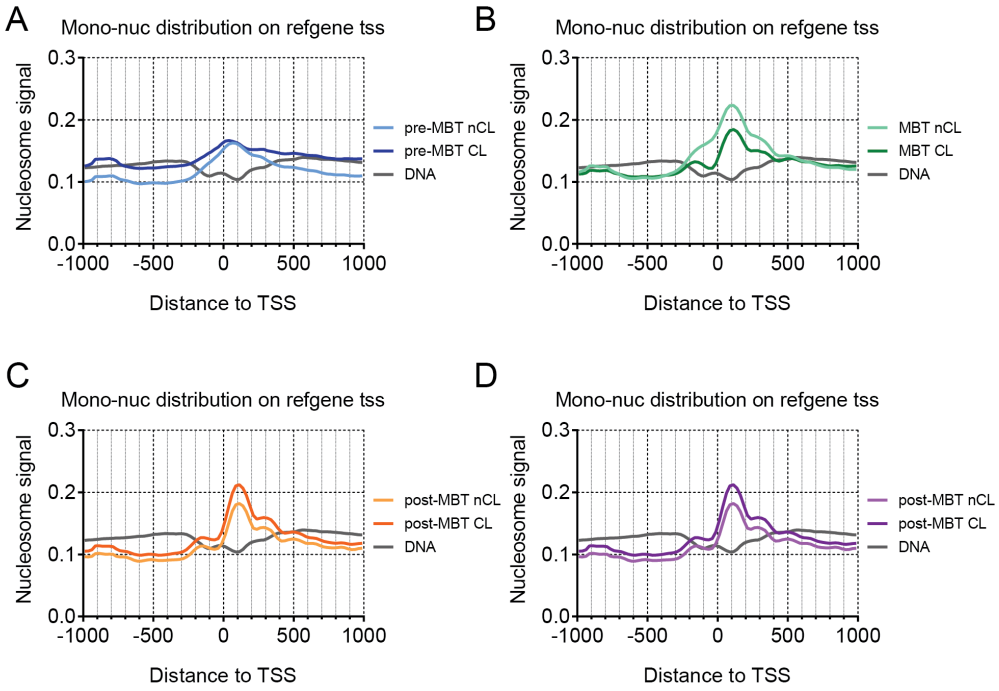


Figure 4.5: Nucleosome distribution of crosslinked and noncrosslinked samples at promoters at each stage. (A-D) The mean profiles of nucleosome signal in CL and nCL samples around TSS of all annotated RefSeq genes at pre-MBT (A), MBT (B), post-MBT (C) and Liver (D).

Figure 4.6: Genome-wide correlation between samples at different stages. (A-C) Pairwise comparisons of normalized nucleosome signal per 100bp in CL samples between pre-MBT and MBT (A), MBT and post-MBT (B), post-MBT and liver (C). (D-F) Pairwise comparisons of normalized nucleosome signal per 100bp in nCL samples between pre-MBT and MBT (D), MBT and post-MBT (E), post-MBT and liver (F). R value represents Pearson Correleation score. (G) Pearson Correlation between any of the 8 samples, clustered into heatmap.

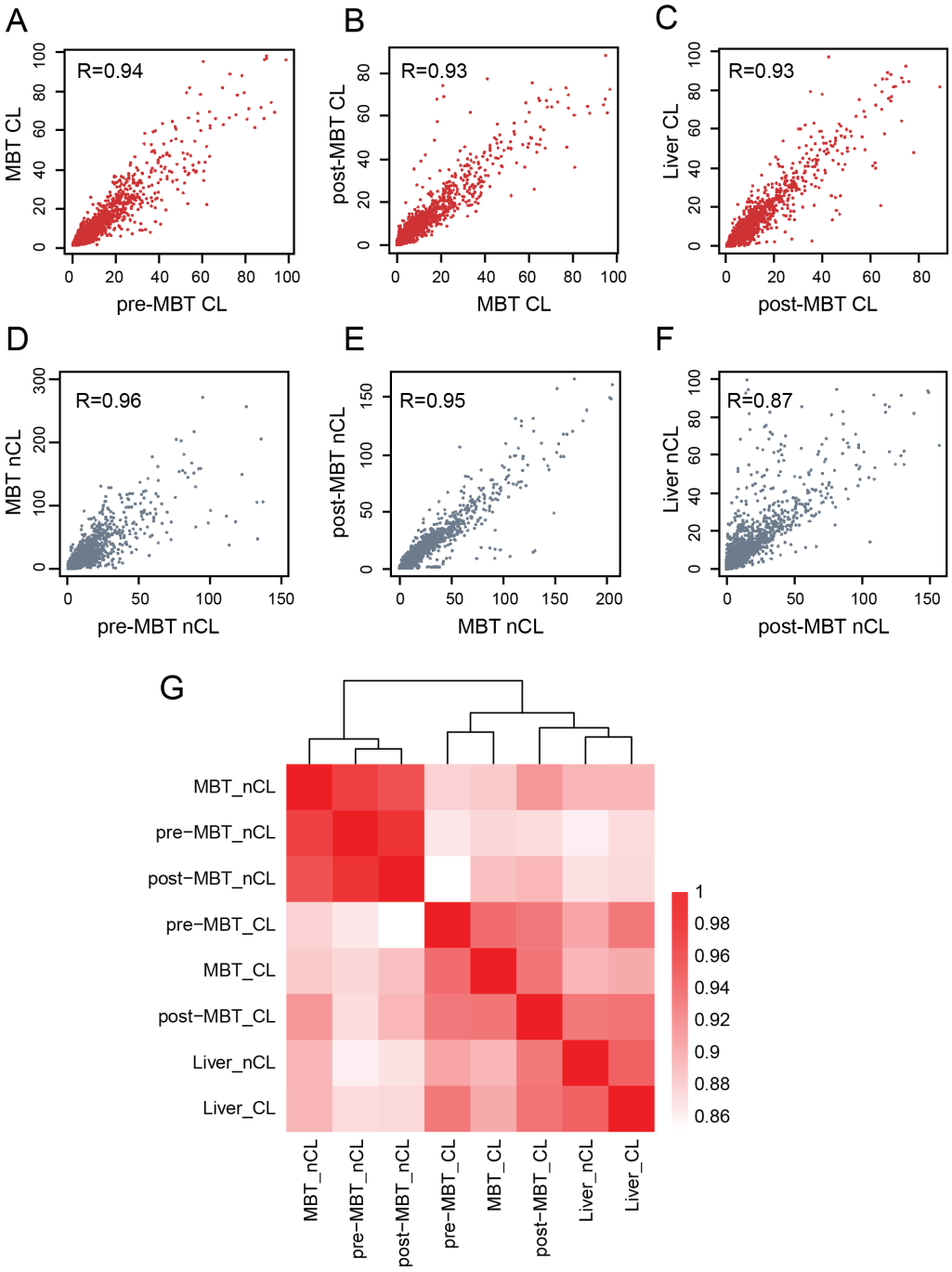
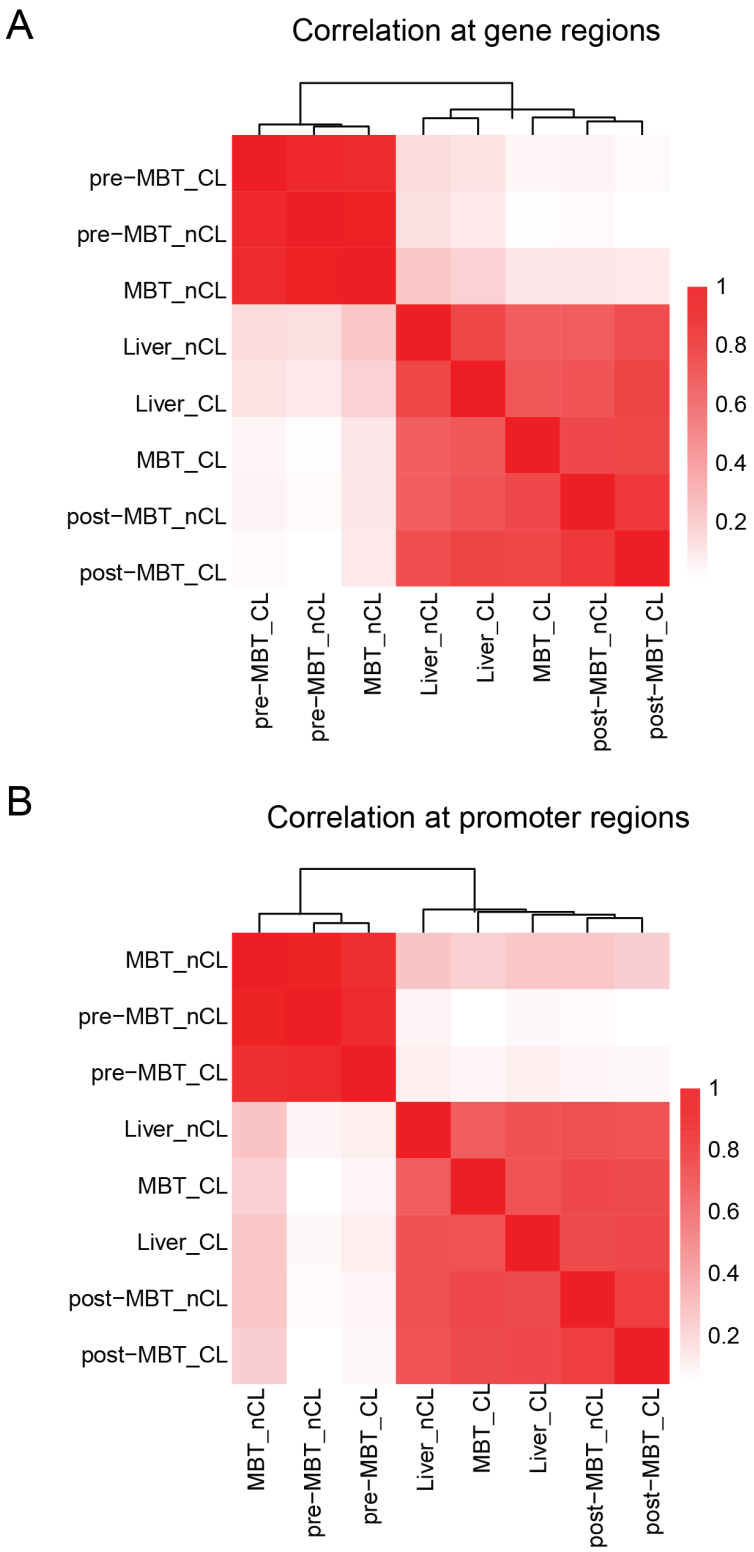

A
R=0.94
MBT CL
pre-MBT CL
B
R=0.93
post-MBT CL
MBT CL
C
R=0.93
Liver CL
post-MBT CL
D
R=0.96
MBT nCL
pre-MBT nCL
E
R=0.95
post-MBT nCL
MBT nCL
F
R=0.87
Liver nCL
post-MBT nCL
G
MBT_nCL
pre−MBT_nCL
post−MBT_nCL
pre−MBT_CL
MBT_CL
post−MBT_CL
Liver_nCL
Liver_CL
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86

Figure 4.7: Correlation heatmaps of all samples at genes and promoters. (A) Pearson Correlation between any of the 8 samples at gene regions, clustered into heatmap. Gene region is defined as areas expanding from upstream 5kb of gene start to downstream 5kb of gene end. Normalized nucleosome signals of all 8 samples are collected at each gene region. (B) Pearson Correlation between any of the 8 samples at gene regions, clustered into heatmap. Promoter region is defined as $\pm$2kb of transcription start site (TSS). Normalized nucleosome signals of all 8 samples are collected at each promoter region.


A
Correlation at gene regions
pre−MBT_CL
pre−MBT_nCL
MBT_nCL
Liver_nCL
Liver_CL
MBT_CL
post−MBT_nCL
post−MBT_CL
1
0.8
0.6
0.4
0.2
B
Correlation at promoter regions
MBT_nCL
pre−MBT_nCL
pre−MBT_CL
Liver_nCL
MBT_CL
Liver_CL
post−MBT_nCL
post−MBT_CL
1
0.8
0.6
0.4
0.2

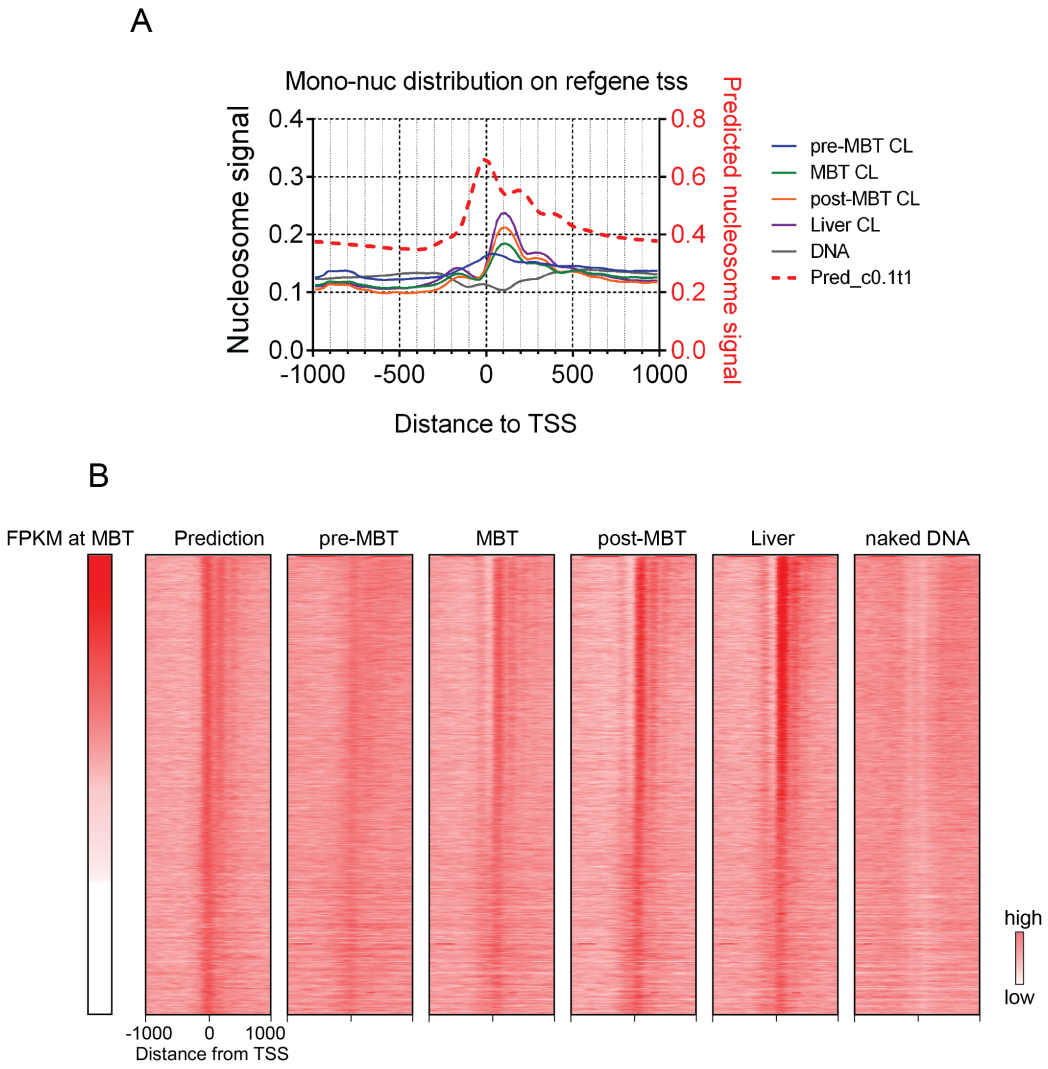


Figure 4.8: Nucleosome array is observed around TSS. (A) Observed nucleosome signals in all CL samples around TSS are averaged and plotted on the left Y axis. Mean value of predicted nucleosome signals are plotted on the right Y axis. (B) Heatmaps of predicted nucleosome signals, observed nucleosome signals at pre-MBT, MBT, post-MBT, liver and naked DNA around TSS. Each row represents $\pm$1kb around TSS, ranked by RNA FPKM value at MBT (4hpf) (leftmost panel).

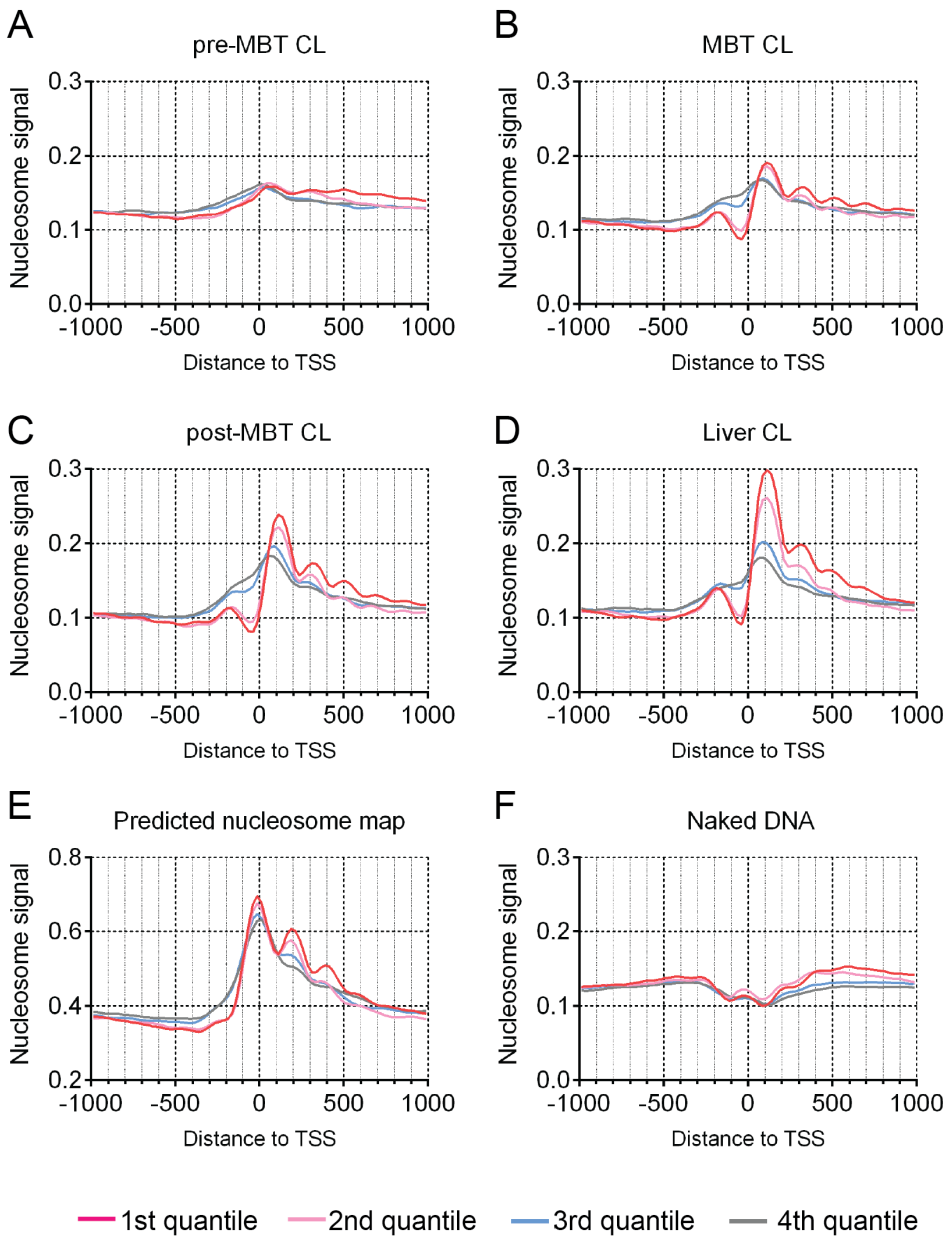


Figure 4.9: Nucleosome array around TSS correlates with expression level. Genes are grouped evenly into 4 quantiles based on the expression level (FPKM) at MBT. First quantile represents the top 25% expressed genes. Fourth quantile represents the bottom 25% expressed genes. (A-F) The mean profiles of nucleosome signal around TSS of four groups genes; nucleosome signals at pre-MBT (A), MBT (B), post-MBT (C), liver (D), predicted nucleosome (E), and naked DNA (F).

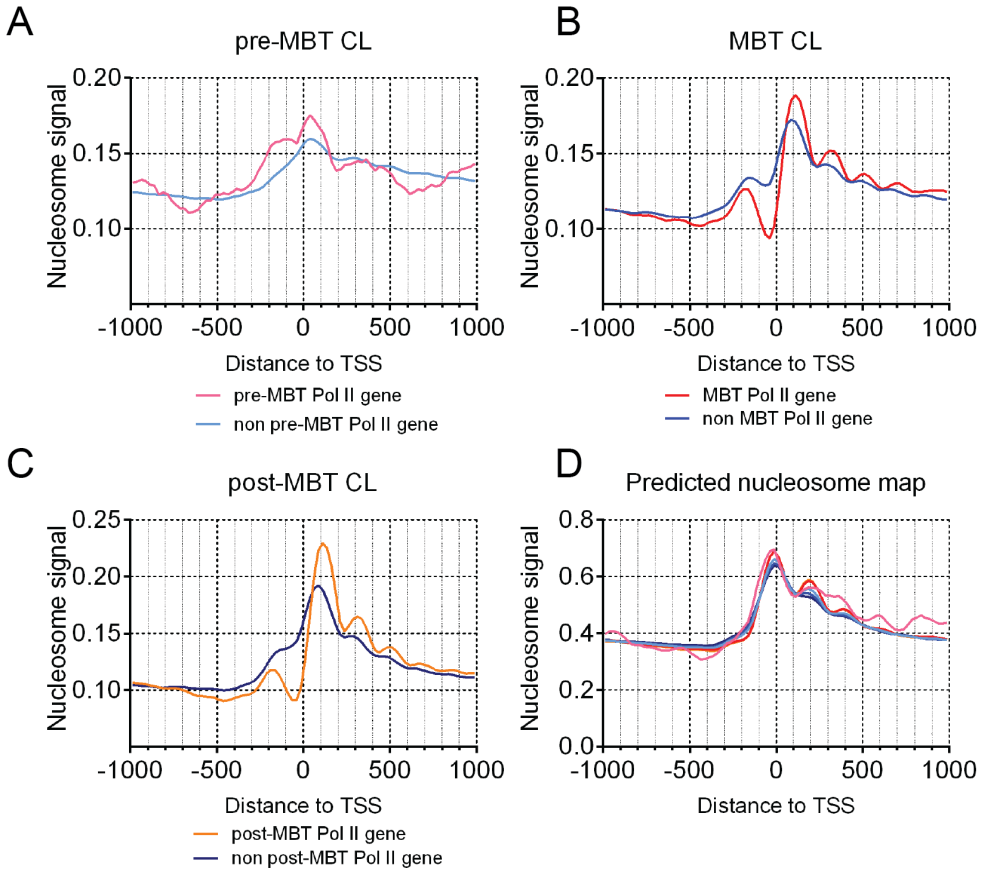


Figure 4.10: Pol II bound genes have strong nucleosome array around TSS. Genes are partitioned into Pol II bound and unbound based on ChIP-seq of Pol II at pre-MBT, MBT and post-MBT. (A) The mean profiles of nucleosome signals at pre-MBT around Pol II bound and unbound genes at pre-MBT. (B) The mean profiles of nucleosome signals at MBT around Pol II bound and unbound genes at MBT. (C) The mean profiles of nucleosome signals at post-MBT around Pol II bound and unbound genes at post-MBT. (D) Predicted nucleosome signals at Pol II bound genes at pre-MBT, MBT and post-MBT, as well as Pol II unbound genes at pre-MBT, MBT, and post-MBT.

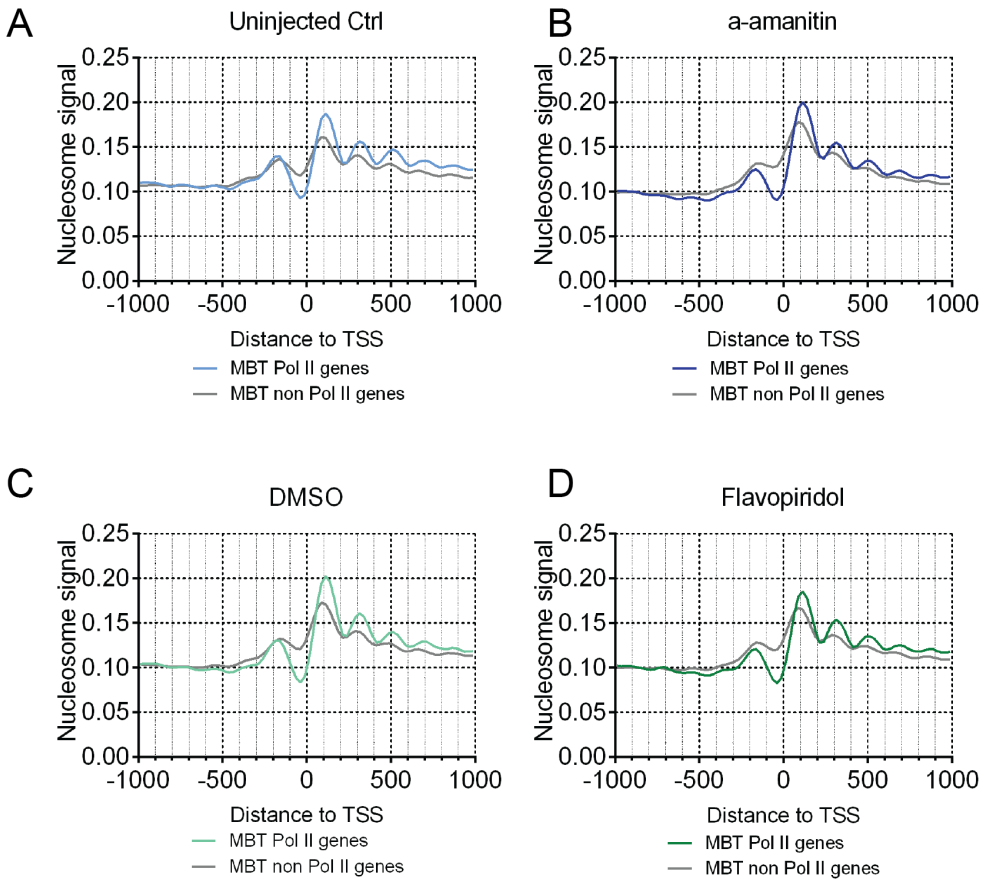


Figure 4.11: Nucleosome array formation is not dependent on transcription. Genes are partitioned into Pol II bound and unbound based on ChIP-seq of Pol II at MBT. The mean profiles of nucleosome signals from uninjected control embryos (A), a-amanitin injected embryos (B), DMSO-treated control embryos (C), flavopiridol-treated embryo (D) around TSS of the two groups of genes.

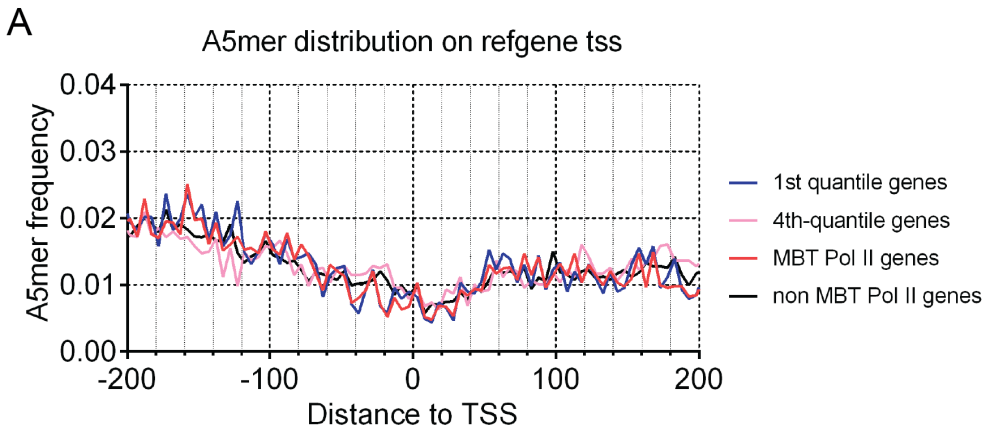


Figure 4.12: Characteristics of DNA sequence around TSS. First quantile: top 25% expressed genes at MBT; Fourth-quantile: bottom 25% expressed genes at MBT. MBT Pol II and non-MBT Pol II: genes bound or unbound by Pol II at MBT. (A) A5mer (AAAAA) sequence frequency around TSS of 4 groups of genes. (B-E) A/T dinucleotide and G/C dinucleotide frequency around TSS of 1st-quantile genes (B), 4th-quantile genes (C), MBT Pol II genes (D) and non MBT Pol II genes (E).

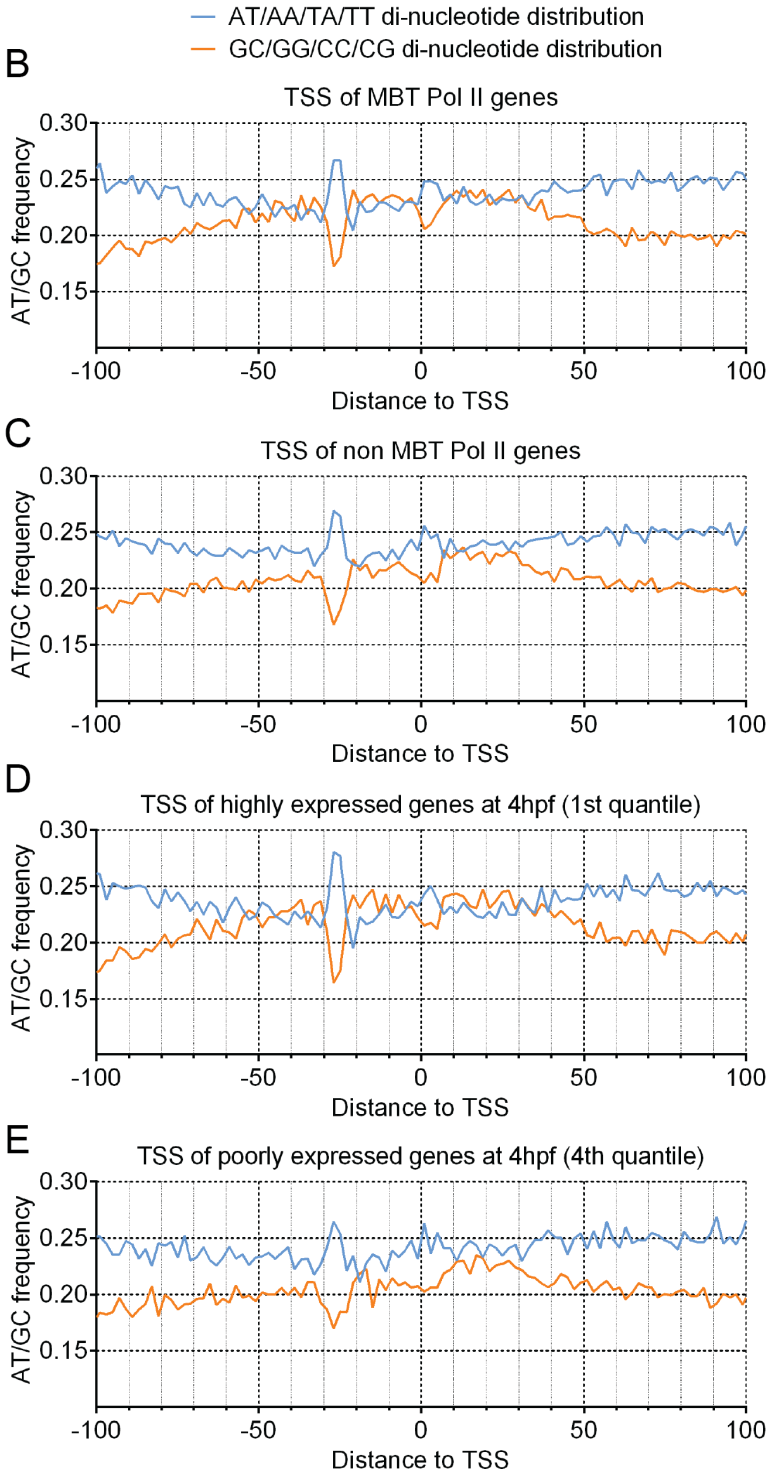


Figure 4.12: Continued.

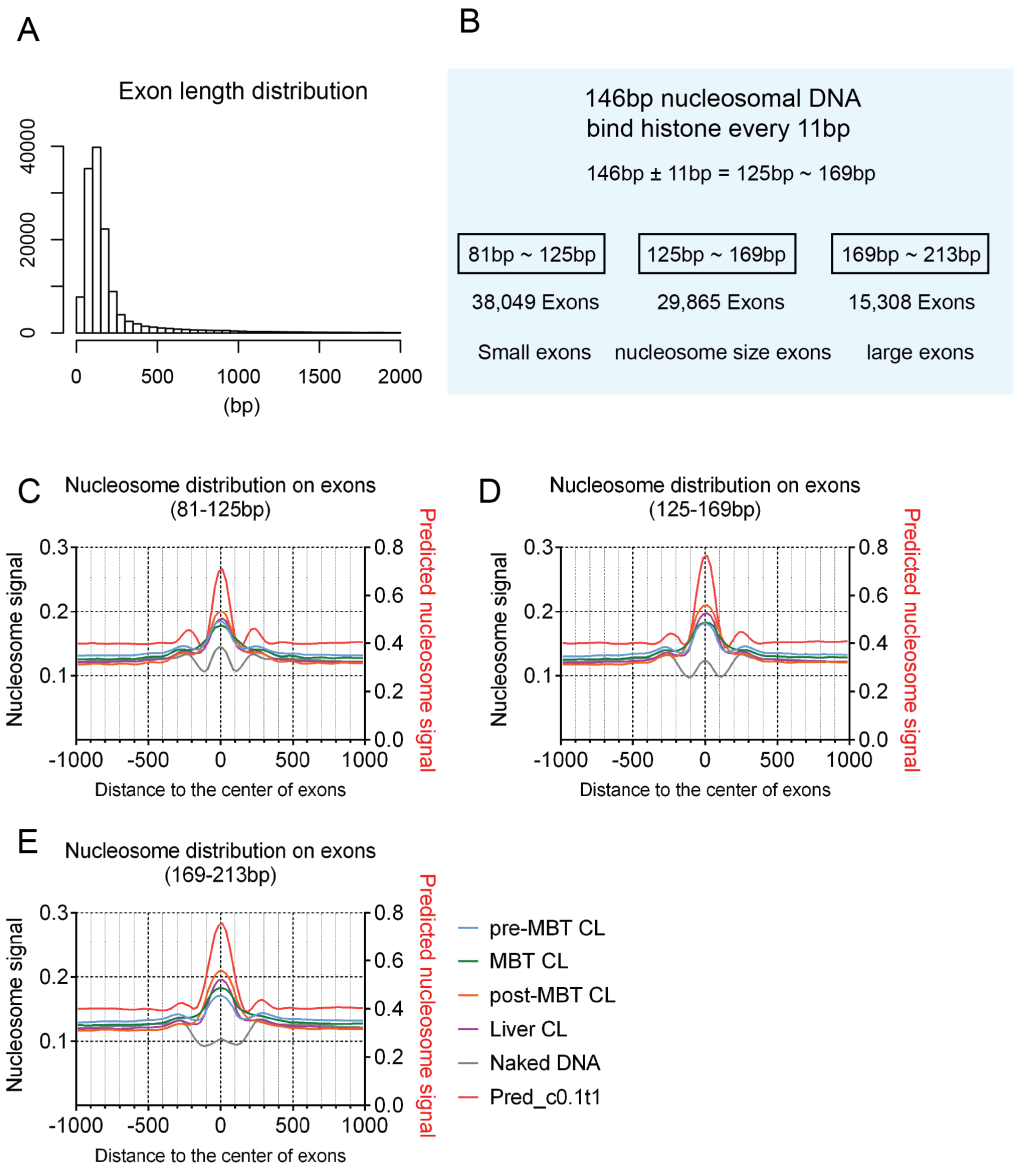


Figure 4.13: Nucleosomes are also enriched on exons. (A) Histogram of exons length of all RefSeq genes. (B) Schematic of separation of exons into 3 groups based on exon length. (C-E) The mean profiles of observed nucleosome signal of all stages and predicted nucleosome signals around the center of 81-125bp exons (C), 125-169bp exons (D), 169-213bp exons (E). Observed nucleosome signal from pre-MBT CL, MBT CL, post-MBT CL, liver CL and naked DNA are plotted on the left Y axis. Predicted nucleosome signal is plotted on the right Y axis.

Figure 4.14: Strong A5mer presents specifically at the 5'end of exons. The mean profiles of A5mer frequency around the 5'end (black) or 3'end of exons (pink). Frequency is collected using 5bp windows expanding $\pm$200bp around either end of 81-125bp exons (A), 125-169bp exons (B); and using 5bp windows expanding $\pm$250bp around either end of 169-213bp exons (C).

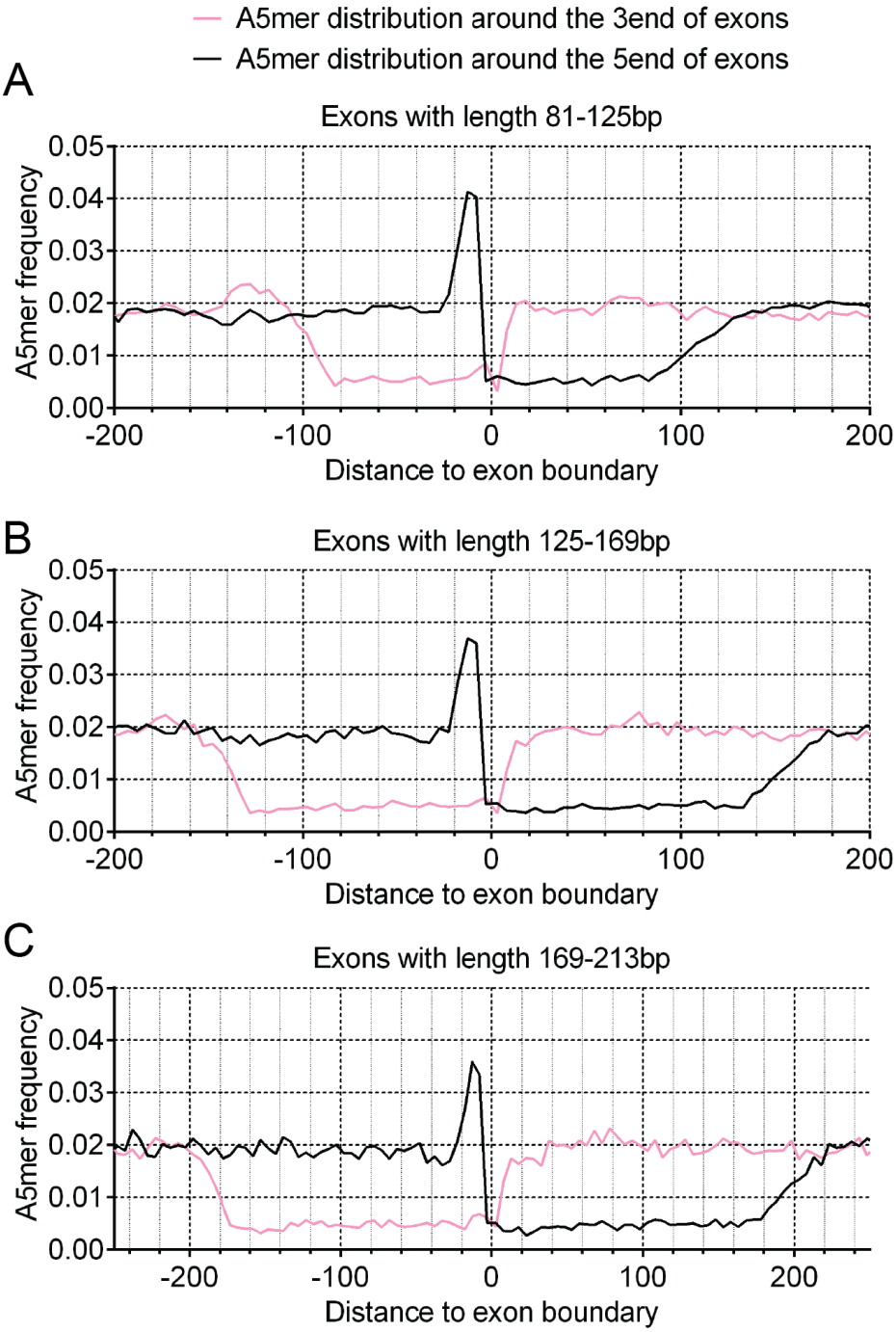

A5mer distribution around the 3end of exons
A5mer distribution around the 5end of exons
A
Exons with length 81-125bp
A5mer frequency
0.05
0.04
0.03
0.02
0.01
0.00
-200
-100
0
100
200
Distance to exon boundary
B
Exons with length 125-169bp
A5mer frequency
0.05
0.04
0.03
0.02
0.01
0.00
-200
-100
0
100
200
Distance to exon boundary
C
Exons with length 169-213bp
A5mer frequency
0.05
0.04
0.03
0.02
0.01
0.00
-200
-100
0
100
200
Distance to exon boundary

Figure 4.15: AT and GC alternation is observed on exons. The mean profiles of A/T(blue) and G/C (orange) dinucleotide frequency around the 5'end of exons. Frequency is collected using 2bp windows expanding $\pm$200bp around the 5'end of 81-125bp exons (A), 125-169bp exons (B); and using 2bp windows expanding $\pm$250bp around the 5'end of 169-213bp exons (C).

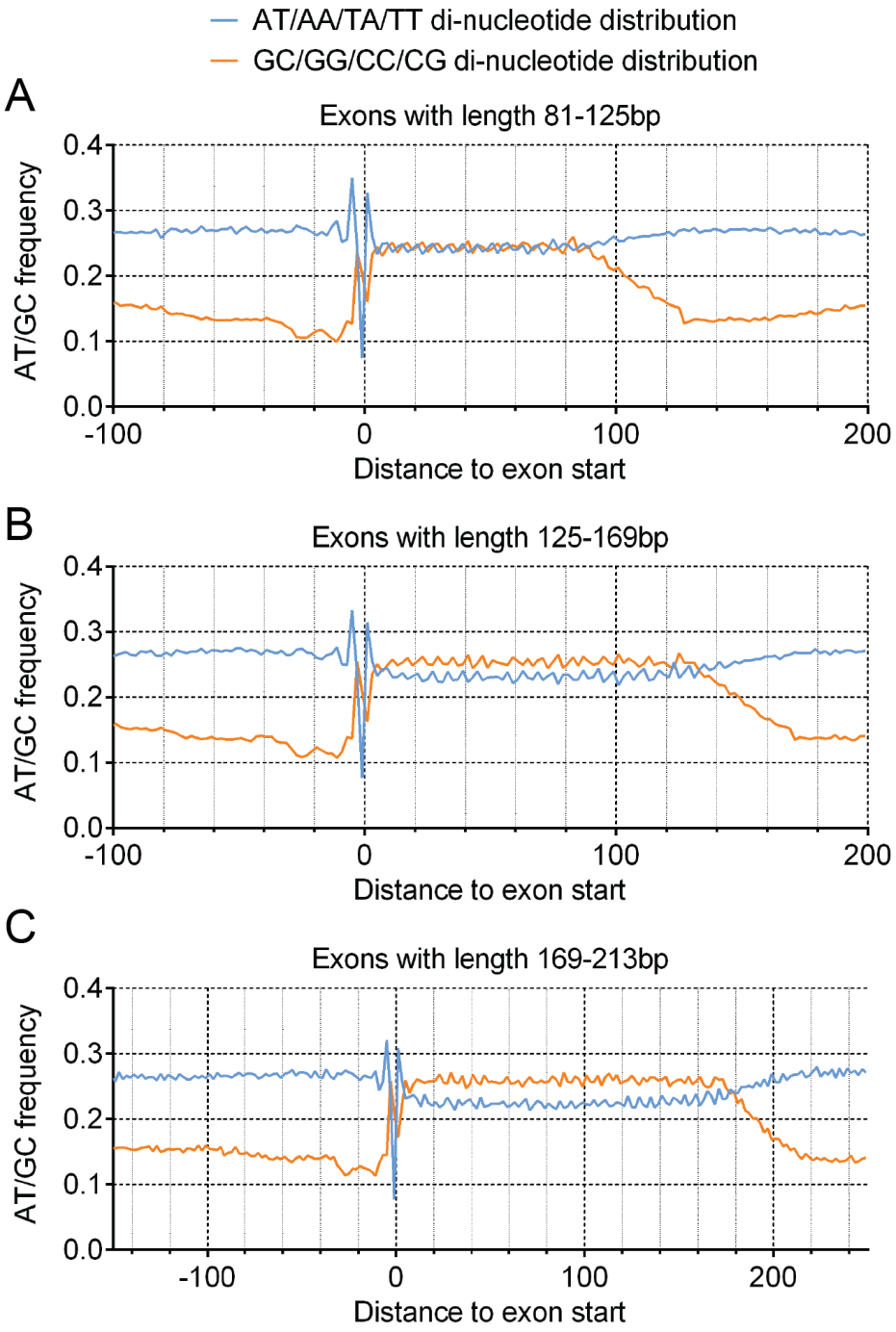

AT/AA/TA/TT di-nucleotide distribution
GC/GG/CC/CG di-nucleotide distribution
A
Exons with length 81-125bp
AT/GC frequency
0.4
0.3
0.2
0.1
0.0
-100
0
100
200
Distance to exon start
B
Exons with length 125-169bp
AT/GC frequency
0.4
0.3
0.2
0.1
0.0
-100
0
100
200
Distance to exon start
C
Exons with length 169-213bp
AT/GC frequency
0.4
0.3
0.2
0.1
0.0
-100
0
100
200
Distance to exon start

Figure 4.16: Clustering of exons based on AT and GC alternation. For each exon, A/T and G/C frequency are counted using 2bp windows around the 5'end of exons. A/T and G/C frequency of each exon are then combined and subject to *k*-mean clustering. (A) 125-169bp exons ($\pm$200bp), (B) 169-213bp exons ($\pm$250bp), (C) 81-125bp exons ($\pm$200bp).

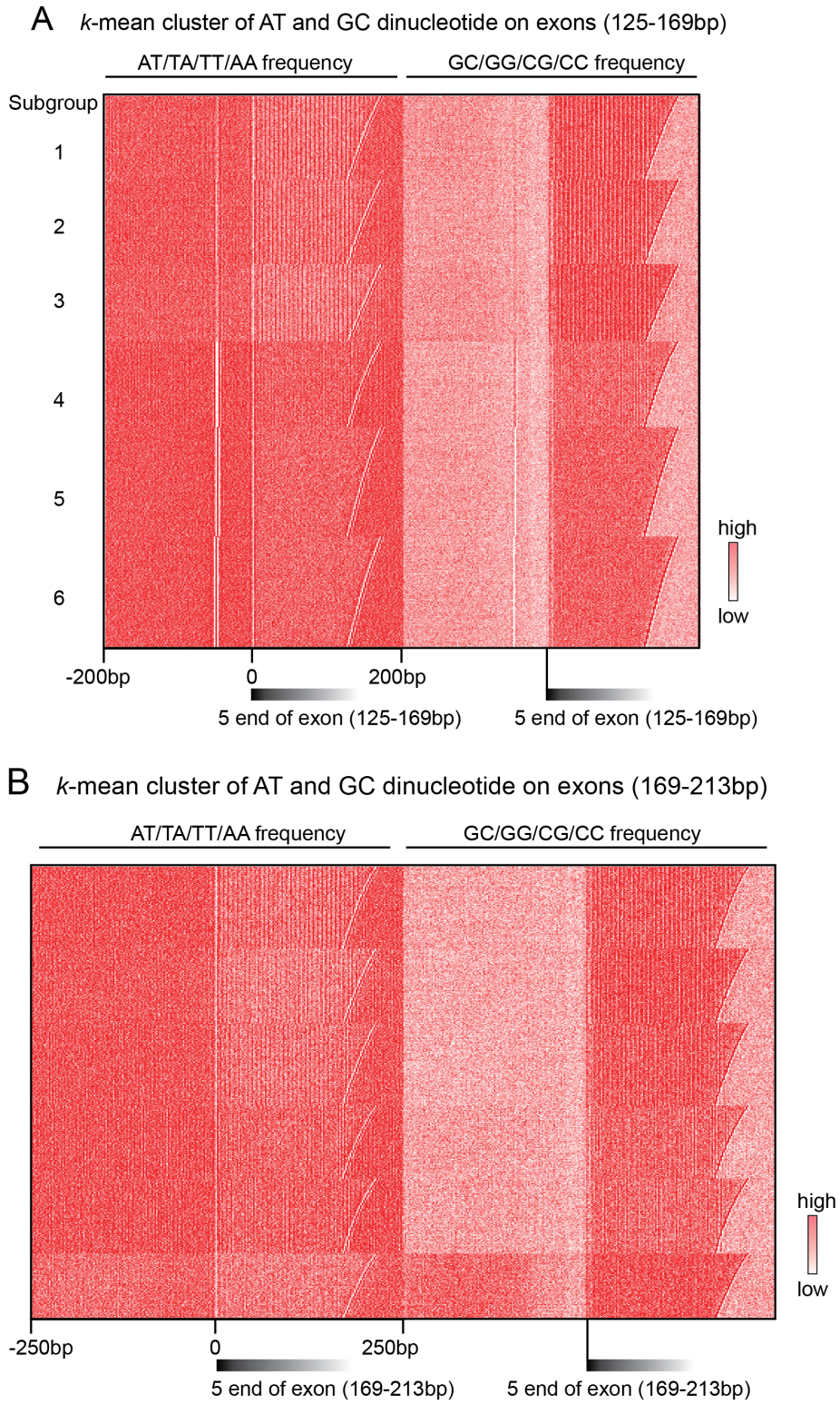


Figure 4.16: Continued.

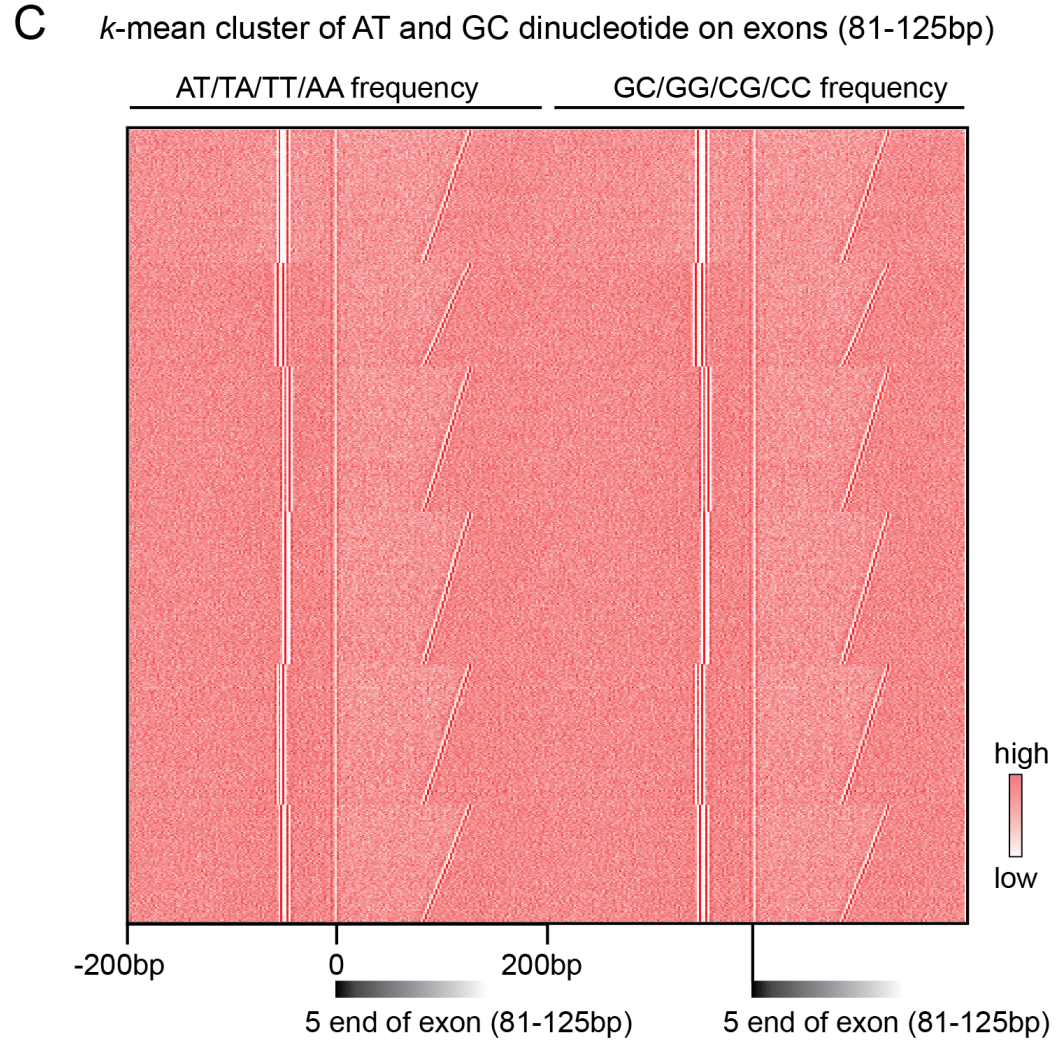
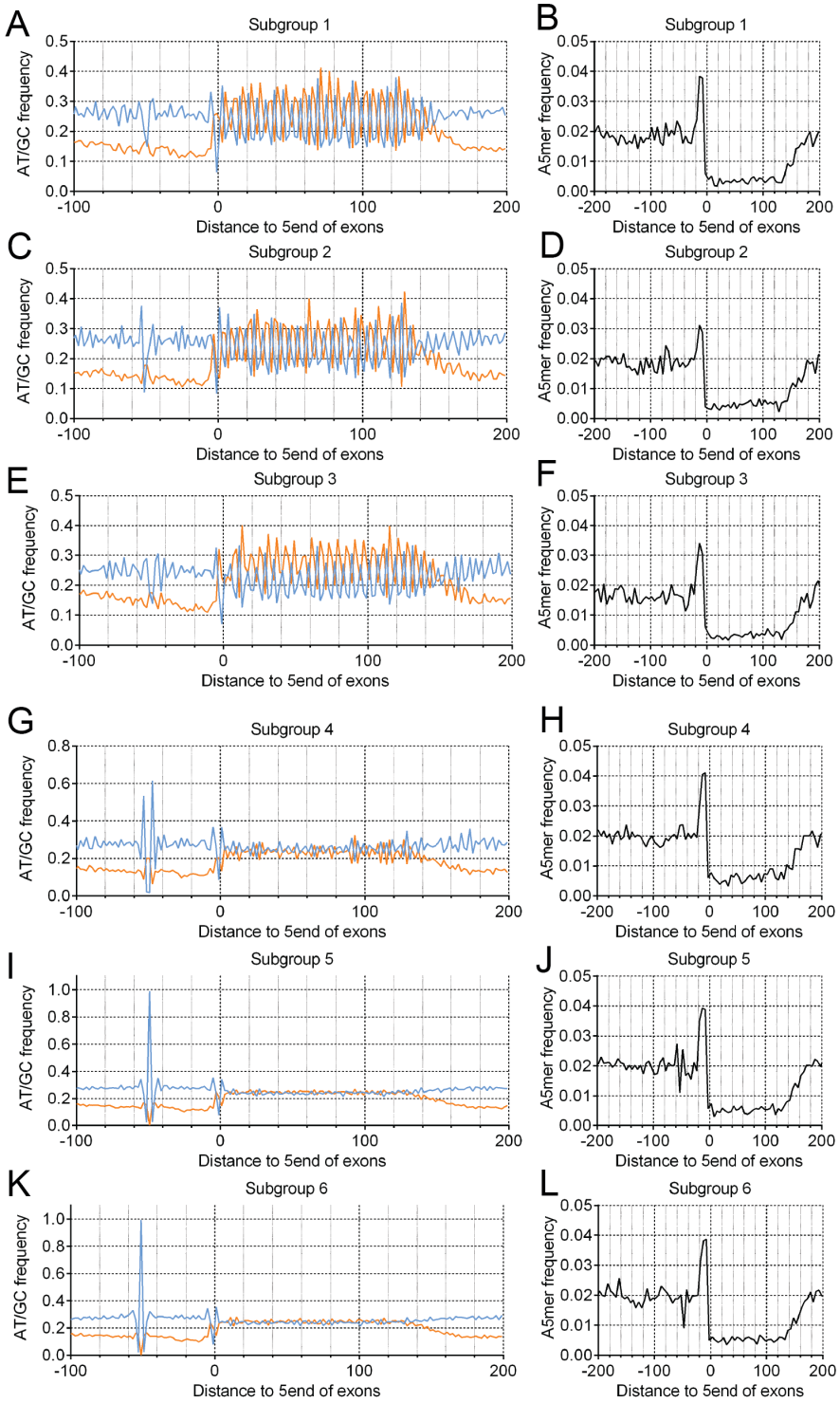


Figure 4.16: Continued.

Figure 4.17: A5mer and AT- GC alternation of each subgroup of 125-169bp exons. The mean profiles of A/T(blue) and G/C (orange) dinucleotide frequency, and A5mer (black) frequency around the 5'end of 125-169bp exons, grouped by *k*-mean clustering. Subgroup 1 (A, B), Subgroup 2 (C, D), Subgroup 3 (E, F), Subgroup 4 (G, H), Subgroup 5 (I, J), Subgroup 6 (K, L).

Zebrafish 125-169bp exons are grouped based on the *k*-mean clustering



Human or Mouse 125-169bp exons are grouped based on the *k*-mean clustering

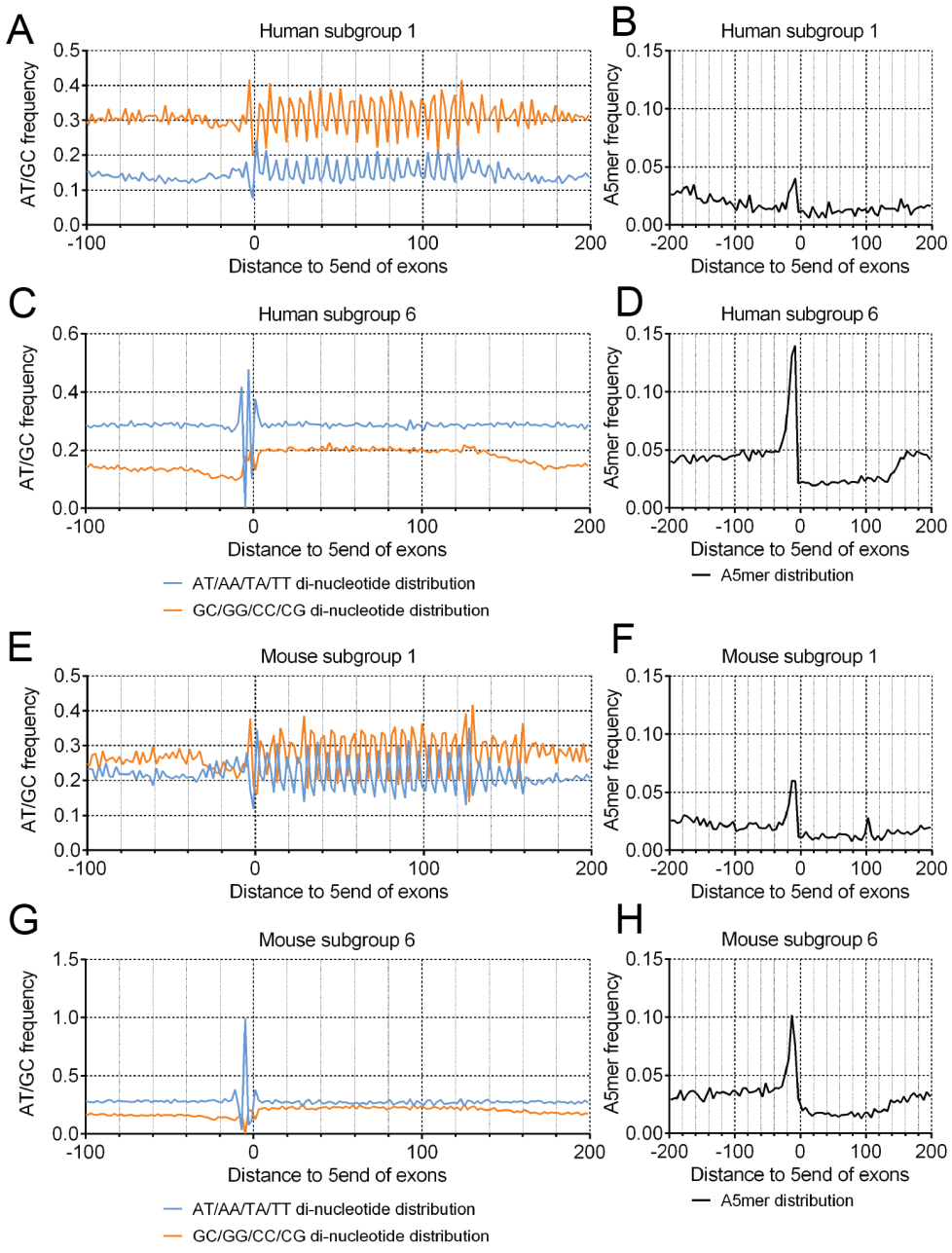


Figure 4.18: AT- GC alternation on exons is conserved in human and mouse. The mean profiles of A/T(blue) and G/C (orange) dinucleotide frequency, and A5mer(black) frequency around the 5'end of 125-169bp exons, grouped by *k*-mean clustering. Subgroup 1 of human exons (A, B), Subgroup 6 of human exons (C, D), Subgroup 1 of mouse exons (E, F), Subgroup 6 of mouse exons (G, H).

CHAPTER 5

SUMMARY AND FUTURE DIRECTION

5.1 Summary of the roles of SWI/SNF complexes in early zebrafish embryo

In Chapter 2, I have described the first survey of genome-wide Brg1 and Brm occupancy in early zebrafish embryos, which allows us to understand how SWI/SNF complexes are involved in early embryogenesis. In addition, our work enables the direct comparison of Brg1 and Brm based on genome-wide occupancy, as there is no genome-wide study of Brm occupancy available to date.

Brg1 and Brm occupancy increased substantially from pre-MBT to MBT. MBT is around when zygotic genome activation (ZGA) happens, where hundreds of genes are activated simultaneously and steeply. The dramatic increase of Brg1 and Brm occupancy implies they may play important roles in ZGA. Indeed, pluripotency genes, *pou5f1, nanog, sox* factors are strongly bound by Brg1 and Brm. Other transcription factors (e.g. *foxh1* and *ddx5*) and epigenetic regulators (e.g. *anp32e* and *brg1*) are also bound by Brg1 and Brm. However, genes required later in development, such as *hox* genes and many *tbx* genes, are not bound by Brg1 and Brm. In addition, Brg1 and Brm may also regulate homeostasis, as many housekeeping genes (e.g. *actb1* and *actb2*) are occupied by Brg1 and Brm. Surprisingly, Brg1 and Brm are highly enriched at miR-430 locus in a stage-specific

manner. The potential roles in activating miR-430 transcription suggest Brg1 and Brm are also involved in degradation of maternally provided factors and promoting the maternal-zygotic transition. Collectively, based on occupancy, Brg1 and Brm function in diverse pathways, and may play important roles in regulating early zebrafish embryogenesis.

To determine whether certain transcription factors are involved in recruiting Brg1 and Brm to specific loci, I performed motif analysis and identified known and unknown motifs at Brg1- and Brm-occupied regions, among which Eomesa, Sox2, Pou5f3 and Nanog are of particular interests. Using published ChIP-seq datasets of these factors, I confirmed they are enriched at Brg1 and Brm binding sites. Moreover, Eomesa, Sox2, pou5f3 and Nanog co-localize with Brg1 and Brm at many key development genes, including *vent, vox, her5*, and themselves (*pou5f1, sox2* and *nanog*).

Comparing their occupancy, we discovered shared and specific roles of Brg1 and Brm in early zebrafish development. First, Brg1 and Brm co-occupy a large number of genes (57 at pre-MBT, 1836 at MBT, and 2854 at post-MBT), but also bound specific genes, respectively. Common genes (co-bound by Brg1 and Brm) are highly transcribed and are enriched for housekeeping functions based on Gene Ontology analysis. In contrast, Brg1-bound specific genes are highly enriched for transcription and development regulation. Brm-bound specific genes were enriched for both housekeeping and development functions. These results suggest Brg1 and Brm co-regulate housekeeping genes, and also have specific functions in regulating development genes.

One feature unique to Brg1 but not to Brm is the high correlation with active

histone marks. Common (co-bound by Brg1 and Brm) and Brg1-specific regions are enriched for active histone marks (H3K27ac, H3K4me1, and H3K4me3), but not repressive mark, H3K27me3, while Brm-specific regions have a considerably lower enrichment for active histone modifications. Another striking observation is the distribution of Brm over gene bodies of around 1000 genes. Although a few studies have reported Brm binding at coding regions of a handful genes (by ChIP-qPCR), it has not been investigated in a genome-wide scale. Our study provides a comprehensive view and reveals Brm occupancy at gene body is prevalent and is in distinct difference with Brg1, which specifically resides in promoter-proximal regions. Importantly, Brm and Pol II occupancy overlap precisely at the gene body, where they almost always start and stop at the same loci. Furthermore, Brm occupancy at the gene body is lost significantly when Pol II elongation is inhibited, which suggests Brm traveling to the gene body is dependent on transcription elongation. Taken together, we demonstrate that Brg1 and Brm may play important roles in early zebrafish development. We also reveal specific roles of Brg1 and Brm, and especially the unknown role of Brm in elongation regulation.

5.2 Summary of the accessible chromatin in early zebrafish embryo

To gain a comprehensive view of the dynamic chromatin profiles during early embryogenesis, ATAC-seq was performed in early zebrafish embryos and adult liver. Total reads were partitioned into ATAC-open and ATAC-mono based on length. ATAC-open data are fragments with a length of 0-100bp and represent open chromatin, whereas ATAC-mono data are fragments with a length of 120-220bp and represent mono-nucleosome. By MACS2 peaks calling and subsequent filtra-

tion with stringent cutoff, I identified 850, 24331, 32934, and 26799 open chromatin regions at pre-MBT, MBT, post-MBT and liver, based on ATAC-open data. The increase of chromatin accessibility is substantial from pre-MBT to MBT, which suggests a generally permissive chromatin is underlying the genome activation. The open chromatin regions at MBT, post-MBT and liver were highly enriched at promoters, and were associated with genes enriched at different biological processes. At MBT, open chromatin regions are enriched for both transcriptional and metabolic processes. At post-MBT, newly gained open chromatin regions are related to development processes including pattern specification and embryonic morphogenesis. As expected, open chromatin regions in liver are enriched for enzymatic pathways. These data suggest our ATAC-seq at each stage is indeed stage specific, and allow us to understand the specific regulatory network at each stage.

At MBT, post-MBT and liver, more than 8000 genes are associated with open chromatin regions, and these genes are expressed at a higher level compared to genes that are not associated with open chromatin regions. In addition, highly transcribed genes have stronger open chromatin, suggesting open chromatin play an important role in gene expression. However, $\sim$16% of genes have open chromatin but are not transcribed. Interestingly, the open chromatin regions of these genes are enriched for H3K4me1 and H3K27me3. This bivalent mark is different from the canonical bivalent marks that are composed of H3K4me3 and H3K27me3. However, I speculate their functions are analogous, which is to poise genes for transcription.

Here, I also explored the open chromatin of Pol III genes. Contrary to the

genome-wide increase of chromatin accessibility from pre-MBT to MBT, Pol III genes gradually lose chromatin accessibility around the TSS from pre-MBT to MBT. Interestingly, active histone modifications are enriched at the promoters of open Pol III genes. In addition, $\sim$16% of Pol III genes reside within 5kb of Pol II genes, and Pol II is enriched at the Open Pol III genes. These observations suggest Open Pol III genes are also associated with active histone modifications and Pol III genes may take advantage of the accessible chromatin established by Pol II genes.

A number of over-represented motifs are discovered in open chromatin regions at different stages, providing candidate transcription factors for future work. Some factors with known roles in early embryogenesis including Pou5f3, SOX factors, are identified at MBT and post-MBT, but not in liver. Other novel factors including Nr2e3 and NFYB have not been implicated in early zebrafish development, and will be interesting candidates for future studies. In line with the motif analysis, Pou5f3 and Sox2, as well as two other factors Eomesa and Nanog, are enriched in the open chromatin regions. They also display a different occupancy pattern at intergenic and promoter regions. In addition, these factors bind both open and closed chromatin and could function as pioneer factors to open up the local chromatin.

Finally, ATAC-mono data can be used to map nucleosome positioning, but cannot fully recapitulate the nucleosome occupancy map generated by MNase-seq.

5.3 Summary of the nucleosome positioning in early zebrafish embryo

Genomic DNA is wrapped with histones, which form nucleosomes. Nucleosomes are the building blocks for higher-order chromatin, and are also essential

for gene regulation. At regions with active transcription, DNA is accessible and can be detected by ATAC-seq. While most of the genome are assembled into nucleosomes, which can be better assessed by MNase-seq. In short, ATAC-seq and MNase-seq are complementary in profiling chromatin landscape, but have their respective advantages.

Therefore, as a complementary to ATAC-seq data, and to gain a comprehensive view of dynamic nucleosome positioning, MNase-seq was carried out in early zebrafish embryos and adult liver. To truly reflect the endogenous nucleosome positioning, MNase-seq experiments were performed at both crosslinked and non-crosslinked conditions. Globally, nucleosome positioning under the two conditions were largely similar, while at genes and promoters, the variation emerged at the weak nucleosome (-1 nucleosome, upstream of TSS). At MBT, crosslinked sample clearly captured the -1 nucleosome, which was missing in the noncrosslinked samples. This is the first illustration of how crosslinking affects nucleosome mapping, and is informative to the establishment of a standard procedure for MNase-seq.

To determine the dynamics of nucleosome positioning during early embryogenesis, nucleosome density across different stages was compared at different levels. At genome-wide scale, nucleosome positioning at various stages was largely similar (Pearson correlation $>$ 0.8), which is not surprising since only less than 2% of the genome are protein coding genes. At the levels of genes and promoters, nucleosome positioning at pre-MBT displayed a significant difference from that of other stages, which might be owing to the absence of transcription at pre-MBT.

Nucleosome array is defined by the co-occurrence of -1 nucleosome upstream

of TSS, nucleosome depletion regions, +1 nucleosome downstream TSS, and the subsequent downstream nucleosomes after +1 nucleosome. I observed highly transcribed or Pol-II bound genes formed nucleosome array around the TSS. Surprisingly, the predicted nucleosome signal also displayed nucleosome array around TSS, but the +1 nucleosome resided right at the center of TSS. The difference between the observed and predicted nucleosome positioning suggests factors other than DNA sequence also influence the establishment of nucleosome. However, the factor was not the actual transcription activities, as nucleosome positioning was almost the same between untreated embryos and embryos treated with a-amanitin and flavopiridol (transcription inhibitors). Instead, our data suggest Pol II occupancy could serve as a barrier and mediate the nucleosome array formation. In addition, Pol II bound genes or highly transcribed genes have a better periodicity of A/T and G/C dinucleotides compared to genes not bound by Pol II or genes with low-level transcription. Furthermore, promoters of zebrafish genes do not have enrichment of polyA sequence, though it is prevalent at the promoters of yeast genes.

In addition, observed nucleosome at all stages and predicted nucleosome were enriched at exons. This enrichment is probably mediated by higher polyA (A5mer) sequence at the 5'end of exons, higher GC content and periodicity of A/T and G/C dinucleotides inside the exons. By *k*-mean clustering, exons of length 125-213bp can be divided into two groups (exons with length larger than 213bp have not been investigated). One group has weak A5mer frequency at the 5'end of exons and strong A/T and G/C dinucleotides alternation inside exons. The other group has strong A5mer frequency at the 5'end of exons and weak A/T and G/C dinucleotides

alternation inside exons. Two groups of exons with different sequence attributes can also be found in other organisms, suggesting these sequence characteristics of exons are conserved.

5.4 Future direction for understanding the roles of SWI/SNF complexes

My work on the genome-wide profiling of Brg1 and Brm in early zebrafish embryos supports the importance of Brg1 and Brm in early embryogenesis, as reported in mouse and fly. Furthermore, distinct roles of Brg1 and Brm are revealed based on our results. However, there are some aspects of zebrafish SWI/SNF complexes that need to be further explored.

SWI/SNF complexes are highly conserved and have been identified in many organisms including yeast, fly, mouse, human (Winston & Carlson, 1992; Cairns et al., 1994; Peterson & Tamkun, 1995). It is unlikely that zebrafish does not have these complexes, especially all conserved subunits of SWI/SNF complexes are present in fish. Moreover, the protein sequences of zebrafish SWI/SNF complexes subunits are highly similar to their human homologs (Table 2.1). Therefore, I believe Brg1 and Brm, the core enzymatic subunits of SWI/SNF complexes, can be used to represent zebrafish SWI/SNF complex. Nevertheless, it would be interesting to know the subunit composition of zebrafish SWI/SNF complexes, and whether there is an embryo specific complex in zebrafish. Purifying the homogenate SWI/SNF complex and verifying each subunit by Mass Spectrometry will address these questions.

Our current work in this study did not have the functional analysis to ascertain the importance of Brg1 and Brm in early zebrafish development. Given an accurate

assessment of their roles in early development, Brg1 and Brm maternal-zygotic null fish are required since both proteins are maternally provided. With the advances of CRISPR technology, it has become feasible to knockout Brg1 and Brm zygotically and maternally. However, this work is still challenging and requires at least 18 months. To date, *yng* is the only *brg1* mutant fish line available, whereas no *brm* null fish has been established (Link et al., 2000). The zygotic *brg1* null embryo is lethal around 6-day postfertilization due to neural defects (Eroglu et al., 2006). To eliminate the maternal provided Brg1 protein, a conditional null in oocytes is required. This involves the establishment of a transgenic line with endogenous *brg1* flanked by loxP sites, and subsequent cross with Zp3-cre fish line (drive the recombination specifically in oocytes). Of note, the success of the above experiment relies on the presumption that Brg1 is not required for oogenesis. As for Brm, we decided to generate the Brm null fish using CRISPR. Four guide RNAs have been designed to target two loci in the ATPase domain of Brm. Cloning and *in vitro* transcription of these guide RNAs were completed. Currently, the purified guide RNAs are ready for injection to validate whether they are capable of disrupting the targeted sequences.

Alternatively, other approaches can be applied to disrupt the functions of the maternal deposited Brg1 and Brm proteins, including morpholino (MO) knock-down, expressing dominant negative constructs and antibody depletion. I have extensively tested MOs to block the translation of Brg1 and Brm, respectively. Unfortunately, Brm MO had minimal impact on the injected embryos, where only mild ventralization defects were discernable even at a very high dose of MO injection. Brg1 MO caused embryo lethality around 5dpf, which is similar to the Brg1 null

fish. However, residual Brg1 is still detectable in the MO injected embryos. In summary, previous attempts using translational blocking MOs fail to deplete the maternally inherited Brg1 and Brm. I also tried to express the dominant negative Brg1 (point mutation K757R inactivates the enzymatic activities of Brg1). However, the dominant negative Brg1 was translated at a low level, which might be inefficient to compete with the wildtype Brg1. Nevertheless, optimizing the polyA tail length, or trying different protein tag (currently tagged with V5) might increase the translational efficiency. As an alternative, a short dominant negative Brg1 that only covers the ATPase domain might be better translated and still be able to antagonize the wildtype Brg1. Finally, injecting Brg1 and Brm antibody may sequester the Brg1 and Brm proteins, which is theoretically possible but has not been tested. Taken together, disrupting the functions of maternally provided Brg1 and Brm is challenging, and previous attempts failed to do so. Alternative approaches are suggested and need to be further explored.

Using Brg1 and Brm as a proxy for SWI/SNF complexes, I observed their occupancy at Pol II genes is robust and prevalent. Could Pol I and Pol III genes also be regulated by SWI/SNF complexes? To test, I intersected the SWI/SNF binding sites with annotated Pol III genes. I identified 221 Pol III genes, most of which are tRNAs, that are highly enriched for Brg1 and Brm. I also observed active histone modifications around the TSS of these Pol III genes. This interesting observation suggests that SWI/SNF complexes could also regulate Pol III gene expression. A followed up analysis on Pol I genes would provide further insights on the gene regulation roles of SWI/SNF complexes.

5.5 Future direction for understanding chromatin accessibility in early zebrafish embryo

The work in Chapter 3 included in-depth bioinformatics analysis for the acquired ATAC-seq data at various stages. There are a few interesting questions that have not been fully investigated by the current analyses or experiments and are discussed as below.

Our ATAC-seq data demonstrated a substantial increase in chromatin accessibility from pre-MBT to MBT, which is around ZGA. The accessible genome is probably required for the 'burst' of genome-wide transcription activation. This observation is consistent with the recent chromatin accessibility profiles in mouse embryos, where the number of accessible regions at early 2-cell stage (before ZGA) is around $\sim$20,000, which increases to $\sim$50,000 at 2-cell stage (ZGA) (Wu et al., 2016). The increase of chromatin accessibility could be required preceding ZGA or might be established due to the transcription at ZGA. Here, it will be interesting to determine how transcription activity shapes the accessibility in early embryos, which can be investigated by combining a-amanitin injection (which inhibits transcription initiation) and ATAC-seq. Interestingly, in mouse embryo, the open chromatin regions before zygotic genome activation (early 2-cell stage in mouse, pre-MBT in zebrafish) are highly enriched at repetitive elements (Wu et al., 2016). Among them, MERVLs are most enriched, which are highly transcribed at ZGA and may activate over 300 genes (Peaston et al., 2004). Since most of my open chromatin regions at pre-MBT are localized at intergenic regions, and a lot of them are close to repetitive elements (by manually scanning through the genome browser), it would be interesting to investigate whether repetitive elements

in zebrafish are also relatively open before ZGA. Notably, only unique reads are aligned for all the ATAC-seq paried-end data. This procedure will naturally filter out the reads generated from repetitive regions. Hence, alternative alignment method and downstream analysis can be performed to test whether repetitive elements are also relatively open in early zebrafish embryo.

Comparing ChIP-seq data of Brg1 and Brm to ATAC-seq data, a large percentage of open chromatin regions are not occupied by Brg1 and Brm. It is known that other chromatin remodelers are involved in the regulation of early development as well. Knockout Mbd3 (subunits of the NuRD complexes) in mice lead to embryonic lethality around implantation (Kaji et al., 2006, 2007). Deletions of ISWI complexes subunits, Batf and Snf2H, are also embryonic lethal (Landry et al., 2008; Stopka & Skoultchi, 2003). In ES cells, members of INO80 (Ep400) and CHD (Chd1, Chd4, Ch6, Chd8) chromatin remodeler family are highly enriched at promoters of active genes (de Dieuleveult et al., 2016). Together, they suggest other factors are also involved in the establishment of accessible chromatin in the early embryo.

In the past, I have generated custom polyclonal antibodies against members of all four families of chromatin remodelers, which are listed in Appendix B. Brg1 and Brm are two of them, which belong to SWI/SNF complexes. I also have Chd1, belongs to CHD complexes; Ino80, belongs to INO80 complexes; Snf2h and Snf2l which belongs to ISWI complexes. Antibodies for Snf2l, and especially Snf2h, are quite promising. However, Chd1 and Ino80 did not recognize their corresponding proteins based on Western blotting use unpurified serum. It is likely that affinity purification will enrich the respective antibodies, and result in specific immunoblotting results. Nevertheless, Snf2h antibodies have been purified, and ChIP-seqs

have been performed at MBT and post-MBT stages, with two replicates for each stage. A close examination of these available data and further investigation of other chromatin remodelers should be followed.

Recent ATAC-seq in mouse embryo suggests open chromatin is evident at promoter with low CG density (Wu et al., 2016). However, others suggest CpG islands in mammals are largely depleted of nucleosome (Wang et al., 2012). It will be intriguing to determine the relationship between CpG density and chromatin accessibility in early zebrafish embryos. A related question is how does DNA methylation affect chromatin state? Moreover, whether different promoter (TATA box containing and TATA-less) types maintain a different chromatin state. In conclusion, it would be interesting to determine whether genes employed different strategies to gain accessibility at promoters and enhancers by examining the characteristics of the cis elements (GC density, TATA box, CTCF), histone codes (ChIP-seq of various histone modifications) and DNA methylation status.

5.6 Future direction for understanding nucleosome positioning in early zebrafish embryo

My nucleosome positioning data revealed very interesting sequence attributes at promoters and exons. Some attributes are probably unique to teleost fish, while some are conserved among vertebrates. There are a couple of questions that are not fully addressed in the current study and are intriguing for future investigations.

Besides transcription start sites, transcription stop site (TTS) is another important regulator of gene expression. In yeast, higher frequency of AA/TT sequences has been observed, which leads to array formation upstream of the TTS (Mavrich et al., 2008a). In addition, nucleosome is depleted at the TTS in both yeast and fly

(Kaplan et al., 2009; Zhang et al., 2009; Mavrich et al., 2008b). Furthermore, cleavage and polyadenylation sites, AATAAA, could serve as a poly(dA:dT) sequence that defers nucleosomes. All this evidence suggests TTS might have a lower nucleosome density which is probably regulated partly through DNA sequences.

In yeast, another poly(dA:dT) sequence, ATATA, has very low nucleosome density, at a similar level to AAAAA (A5mer) (Kaplan et al., 2009). Both ATATA and AATAAA (cleavage and polyadenylation sites) are different from the A5mer sequence that I have profiled for TSS and exons. Notably, there are 1024 possible sequence combinations for a 5nt long DNA. It is possible that zebrafish employed poly(dA:dT) sequences other than A5mer. To evaluate which poly(dA:dT) sequence is preferably used in zebrafish, averaged nucleosome density on all 1024 possible combinations of a 5nt sequence (5mer) can be collected and compared. It is also likely that at TSS or TTS, different poly(dA:dT) sequences are enriched, which can be tested by plotting the frequency of these 1024 5mers as the A5mer sequence.

In the first paper proposing the statistical positioning model, the author speculated that sites bound by proteins might serve as barriers. Based on this assumption, both enhancers (usually bound by p300) and transcription factors binding sites could act as boundaries, and will be interesting to test. Enhancers can be defined by ChIP-seq data of H3K4me1 and H3K27ac. Nanog, Pou5f3, Sox2, and Eomesa are important regulator of early embryogenesis and have been proposed in Chapter 2 to serve as pioneer factors. In addition, they are enriched in Brg1- and Brm-bound regions, as well as open chromatin regions determined by ATAC-seq. It would be interesting to know whether their binding could be informative to place

nucleosomes and shape the chromatin landscape.

In the past, I consistently observed a unique type of nucleosome maps which is characterized by low nucleosome density genome-wide and extremely high nucleosome density over exons. This type of nucleosome positioning has been observed in many cell types, including cells from early embryo (2.5hpf, 4hpf and 5.3hpf), cells from 5dpf embryo, liver cells, and ZF4 cells (zebrafish fibroblast cell lines (Driever & Rangini, 1993)). As I already demonstrated in the Results of Chapter 4, exons harboring specific sequences enhance the affinity between exons and nucleosomes. It is possible that nucleosomes are most stable on exons and are preferentially retained, whereas weak nucleosomes are lost under certain experiment conditions. This result, if it is true, suggests endogenous DNA is a key determinant of nucleosomes in zebrafish. However, in other experiment conditions, as what I did in Chapter 4, both strong and weak nucleosomes are equally represented. Therefore, under normal experiment conditions, the dominant role of DNA sequence in positioning nucleosome is impossible to perceive.

To confirm the high enrichment of nucleosome on exons truly reflect a group of most stable nucleosome, I have to be able to replicate this unique nucleosome map, and understand what experiment conditions lead to this result. However, I cannot fully replicate these maps even with exhaustive attempts testing many different aspects of MNase-digestion. Recently, two nucleosome maps were compared; one is from Aug 2014, and the other is from April 2016. They both used the same mono-nucleosome DNA but were generated with different reagents during the steps of library preparation and Illumina sequencing. The two nucleosome maps are largely similar, which rules out the possibility that exonic sequences are over-

represented during the PCR amplification step of library preparation and Illumina sequencing. In addition, preliminary data suggest composition of the nucleosome digestion buffer may influence the chromatin digestion and the corresponding nucleosome map. A closer examination of this issue will help us to ascertain whether DNA sequence plays a dominant role in determining the strongest nucleosomes formation on exons.

5.7 Concluding remarks

During early embryogenesis, the maternally provided factors are gradually degraded while the zygotic genome undergoes 'minor' and 'major' wave of gene activation and reaches a pluripotent state. These embryonic cells are pluripotent because they have the potential to differentiate into any cell lineages. Elucidating how pluripotency is regulated is essential for our understanding of development, reprogramming, regeneration and tumorigenesis.

The remarkable plasticity of pluripotent cells relies on a unique transcriptome. Specifically, genes that are required for early development and basic homeostasis are selectively activated, whereas genes functioning during lineage commitment are poised in a repressed but competent state. This transcription circuit is tightly orchestrated at multiple levels, including chromatin accessibility. Chromatin accessibility controls the binding of transcription factors and coregulators to promoters and enhancers, and thus influences gene expression. In this dissertation, I focused on understanding what is the chromatin accessibility landscape in the early zebrafish embryo, and how that is regulated by SWI/SNF complexes. To answer these questions, I have employed a variety of techniques including biochemistry (for

antibody generation), molecular biology, genomics (ChIP, ATAC-seq and MNase-seq) and bioinformatics. I observed the genome is largely refractory at pre-MBT (before zygotic genome activation) which becomes increasingly accessible at MBT and post-MBT (after zygotic genome activation). Contrary to the genome-wide increased accessibility, chromatin around Pol III genes becomes gradually inaccessible along the early development. Motif analysis of the accessible chromatin regions allows me to identify candidate transcription factors (Nr3e2, Nfyb, Pou5f3, Sox factors, Nanog and Eomesa), which may be involved in the establishment of open chromatin regions and/or the regulation of the unique transcriptome in the early embryo. Chromatin accessibility is modulated by chromatin remodelers, one major class of which are SWI/SNF complexes. My work revealed the important roles of SWI/SNF complexes (represented by the ATPase subunits Brg1 and Brm) in early zebrafish development, as well as specific roles of Brg1 and Brm. Specifically, Brg1- but not Brm-bound regions are highly enriched for active histone modifications, whereas Brm but not Brg1 displayed prevalent distribution over gene bodies. In addition, Brm precisely co-localize with Pol II at the gene body, which is due to the association with Pol II during transcription elongation, suggesting Brm may function in transcription elongation. Additionally, nucleosome positioning data allow me to discover many interesting sequence characteristics on promoters and exons, which provide novel insights on how nucleosomes are placed in the genome.

Many interesting observations have been made in the dissertation, some of which need to be further validated or explored for their biological impact. In particular: whether the identified candidate transcription factors indeed play roles in

early embryogenesis; whether the important roles of SWI/SNF complexes can be validated by functional studies; what is the interplay between transcription factor, histone modification and SWI/SNF complex in shaping the chromatin accessibility; what are the functions of Brm enrichment over gene bodies, and what are the underlying mechanisms for the specialized function of Brg1 and Brm?

5.8 References

Cairns, B.R., Kim, Y.J., Sayre, M.H., Laurent, B.C., Kornberg, R.D. (1994). A multisubunit complex containing the swi1/adr6, swi2/snf2, swi3, snf5, and snf6 gene products isolated from yeast. Proceedings of the National Academy of Sciences *91*, 1950–1954.

de Dieuleveult, M., Yen, K., Hmitou, I., Depaux, A., Boussouar, F., Dargham, D.B., Jounier, S., Humbertclaude, H., Ribierre, F., Baulard, C., et al. (2016). Genome-wide nucleosome specificity and function of chromatin remodellers in es cells. Nature .

Driever, W., Rangini, Z. (1993). Characterization of a cell line derived from zebrafish (brachydanio rerio) embryos. In Vitro Cellular & Developmental Biology-Animal *29*, 749–754.

Eroglu, B., Wang, G., Tu, N., Sun, X., Mivechi, N.F. (2006). Critical role of brg1 member of the swi/snf chromatin remodeling complex during neurogenesis and neural crest induction in zebrafish. Developmental Dynamics *235*, 2722–2735.

Kaji, K., Caballero, I.M., MacLeod, R., Nichols, J., Wilson, V.A., Hendrich, B. (2006). The nurd component mbd3 is required for pluripotency of embryonic stem cells. Nature Cell Biology *8*, 285–292.

Kaji, K., Nichols, J., Hendrich, B. (2007). Mbd3, a component of the nurd co-repressor complex, is required for development of pluripotent cells. Development *134*, 1123–1132.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al. (2009). The dna-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Landry, J., Sharov, A.A., Piao, Y., Sharova, L.V., Xiao, H., Southon, E., Matta, J., Tessarollo, L., Zhang, Y.E., Ko, M.S., et al. (2008). Essential role of chromatin remodeling protein bptf in early mouse embryos and embryonic stem cells. PLoS Genet *4*, e1000241.

Link, B., Fadool, J., Malicki, J., Dowling, J. (2000). The zebrafish young mutation acts non-cell-autonomously to uncouple differentiation from specification for all retinal cells. Development *127*, 2177–2188.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., Pugh, B.F. (2008a). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research *18*, 1073–1083.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., et al. (2008b). Nucleosome organization in the drosophila genome. Nature *453*, 358–362.

Peaston, A.E., Evsikov, A.V., Graber, J.H., De Vries, W.N., Holbrook, A.E., Solter, D., Knowles, B.B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Developmental Cell *7*, 597–606.

Peterson, C.L., Tamkun, J.W. (1995). The swi-snf complex: a chromatin remodeling machine? Trends in Biochemical Sciences *20*, 143–146.

Stopka, T., Skoultchi, A.I. (2003). The iswi atpase snf2h is required for early mouse development. Proceedings of the National Academy of Sciences *100*, 14097–14102.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research *22*, 1798–1812.

Winston, F., Carlson, M. (1992). Yeast snf/swi transcriptional activators and the spt/sin chromatin connection. Trends in Genetics *8*, 387–391.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., et al. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature *534*, 652–657.

Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., Struhl, K. (2009). Intrinsic histone-dna interactions are not the major determinant of nucleosome positions in vivo. Nature Structural & Molecular Biology *16*, 847–852.

APPENDIX A

COMMERCIAL ANTIBODY VALIDATION

Zebrafish is a valuable model system that has been used in many different fields. However, one disadvantage is the limited commerically available antibodies. Here, I summarized all the commercial antibodies that have been tested (Table A.1. and Table A.2.). Most antibodies are against transcription factors, and some are against chromatin remodelers. Antibodies against Pou5f3 and Foxh1 work well for both Western blotting and ChIP. The Western blotting results of promising antibodies are presented in following figures.

Table A.1: Antibodies which recognize the corresponding proteins

Antibody	Source	Company	Cat#	Lot#	WB	ChIP-WB
Pou5f1	pRb	Anaspec	55832 CT	JK3013	1:1000	YES
			55832 CT	JK3014	1:1000	YES
			55833 IN	JK3002	1:1000	NA
Foxh1	pRb	Anaspec	55671 IN	JK3004	1:1000	YES
				JK3003	1:1000	NA
NF-kb2	pRb	Anaspec	55484 IN	JL2804	1:1000	NA
b-catenin	pRb	Sigma	C2206	NA	1:2000	YES
TBP	mMs	millpore	MAB3658	NG1837795	1:1000	YES
Sox2	Goat	R&D	AF2018	KOY0110041	1:1000	NA
PolII	mMs	Covance	MMS-126R	8WG16	1:1000	YES
	mMs	Active Motif	102660	NA	1:1000	YES
Lamin B2	mMs	Abcam	Ab8983	NA	1:1000	NA
Brg1*	mMs	Santa Cruz	sc-17796	D2012	1:2000	NA

* Brg1 antibodies from different lot# have varies results

Table A.2: Antibodies which do not recognize the corresponding proteins

Antibody	Clonal	Company	Cat#	Lot#
Foxh1	pRb	Anaspec	55670 CT	JK3007
				JK3008
Snail1a	pRb	Anaspec	55648 IN	JG3005
Snail1b	pRb	Anaspec	55649 IN	JI0904
Sox11a	pRb	Anaspec	55854 NT	JF2107
Pea3	pRb	Anaspec	55589 CT	JG0107
				JG0108
TP53	pRb	Anaspec	55915 IN	JK0403
beta-catenin 1/2	pRb	Anaspec	55365 CT	JH1211
Hmga1	pRb	Abcam	Ab4078	GR7262-1
Lef1	pRb	Abcam	Ab52017	906861
Smad3	mMs	Cell Signaling	C67h9	7#5
p-smad3-Ser423/425	mMs	Cell Signaling	C25A9	#10
Brg1	pRb	Abcam	Ab4801	NA
Brg1	pRb	Cell Signaling	3508S	A52, #1
Brm	pRb	Abcam	Ab15597	NA
Brm	pRb	Fisher	PA5-34597	NA
BAF155	pRb	Santa Cruz	SC-10756 X	Lot # A0813
BAF170	pRb	Santa Cruz	SC-10757 X	Lot # G1603
CHD7	pRb	Abcam	Ab31824	NA
PolII	pRb	Abcam	817-100	NA

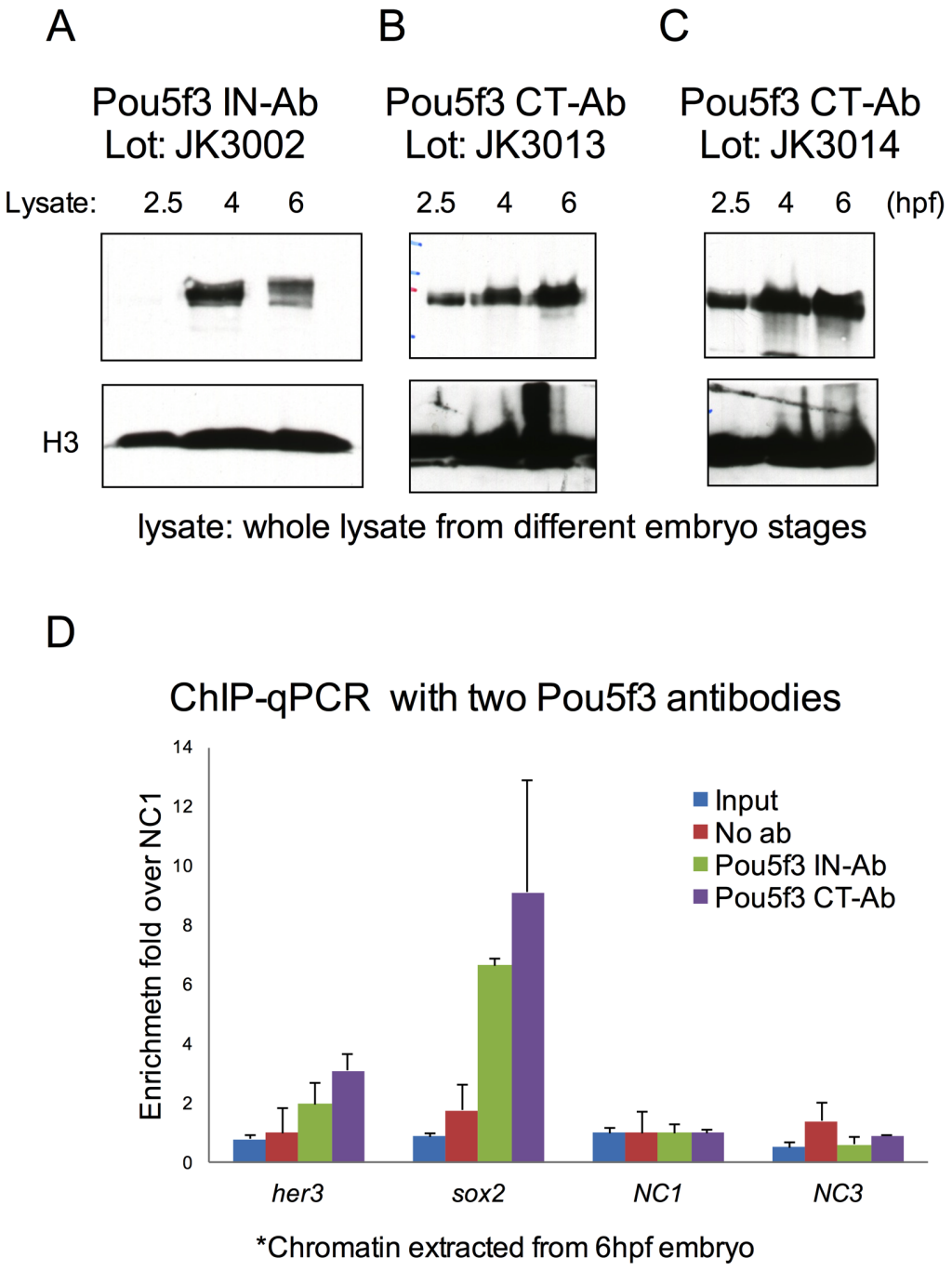


Figure A.1: Western blotting and ChIP-qPCR of Pou5f3. (A-C) Western blotting against Pou5f3 using different antibodies. Whole embryo lysate from different stages are used. (D) ChIP-qPCR of Pou5f3 using antibody recognizing the internal protein sequence (IN) and antibody recognizing the C-terminus protein sequence (CT).

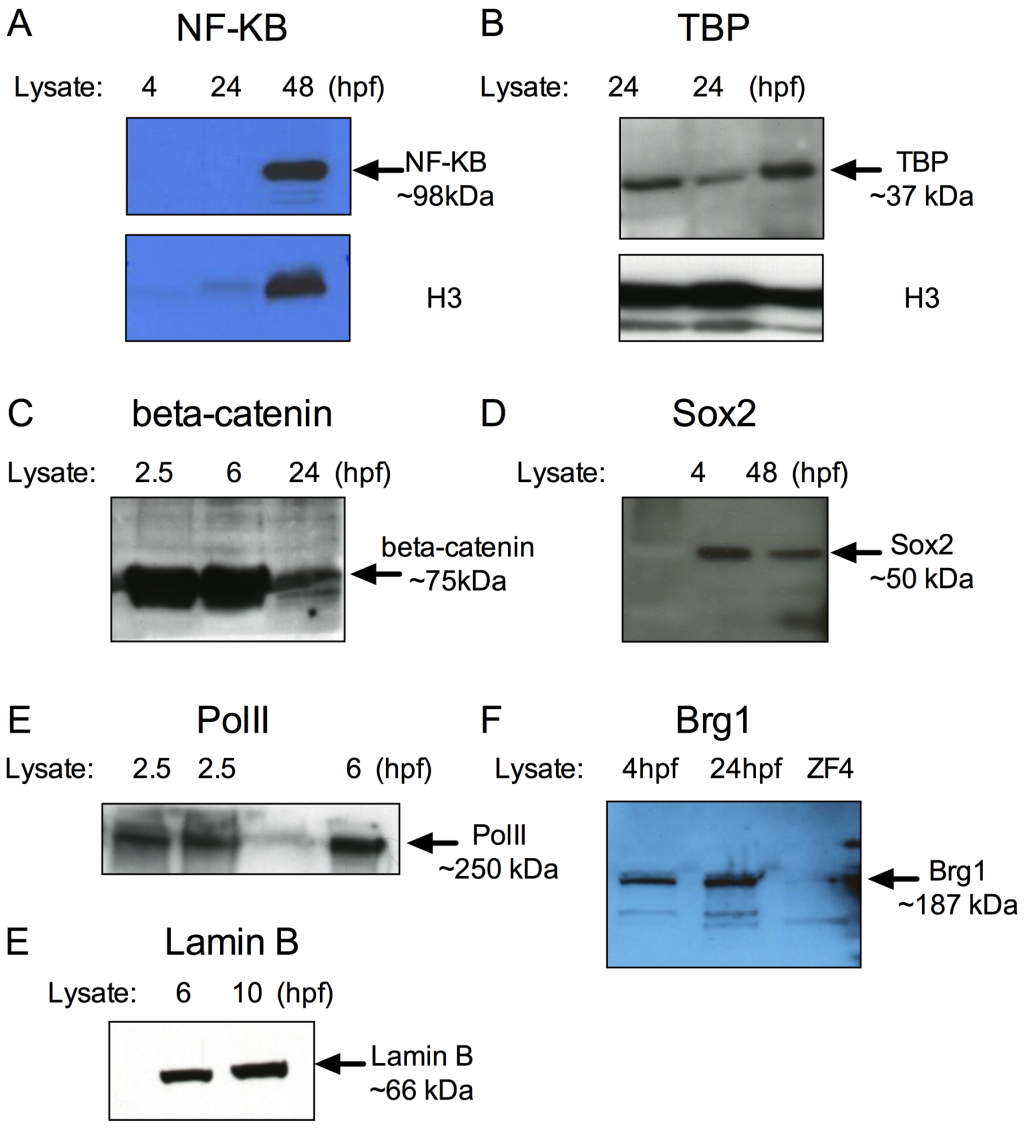


Figure A.2: Validated commercial antibodies. (A-E) Immunoblotting against different protein with their specific antibodies.

APPENDIX B

ADDITIONAL CUSTOMIZED POLYCLONAL ANTIBODIES

Chromatin remodelers can be classified into four major families, including SWI/SNF, ISWI, CHD and INO 80 families. In general, SWI/SNF complexes eject and slide nucleosome to promote transcription activation, while most ISWI family complexes (ACF, CHRAC) assemble nucleosome in a regularly spaced manner to compact nucleosomes and repress transcription. CHD family complexes are involved in both transcription activation (CHD1) and repression (NuRD). INO80 family complexes not only regulate transcription, but also play a role in DNA repair and DNA replication. To gain a comprehensive understanding on the different roles of chromatin complexes in the regulation of early zebrafish development, I have customized polyclonal antibodies against additional chromatin remodeler complexes besides Brg1 and Brm, including BAF180, SNF2H, SNF2L, CHD1 and INO80 (Table B.1.).

Additionally, I am very interested in the roles of transcription factors in the regulation of zygotic genome activation. I first tested many commercial antibodies, and found Pou5f3 and Foxh1 antibodies are very promising. However, to comprehensively investigate the roles of transcription factors, I would like to include some general transcription factors, and a few development regulators. To this end, I have generated polyclonal antibodies against different transcription factors,

including Oct1, Sp1, Sox2, Eomesa and Zeb2b (Table B.1.).

B.1 Peptide expression

The principle for designing antigen is similar to that for Brg1 and Brm, where around 40AA at N-terminal and C-terminal are chose. However, for SNF2H and SNF2L, because C-terminal regions are identical between the two proteins, only the different N-terminal regions are used (Figure B.1). The cloning strategy for the 10 proteins is also similar to that for Brg1 and Brm, except for SNF2L and Sox3. For SNf2L and Sox3, two long oligos coding the entire peptide were synthesized, annealed, and directly ligated to the pET-duet1 vector backbone cut by BamHI and NotI. The long oilgos and primers are listed in Table B.2.

B.2 Antibodies result

As indicated in Table B.1., out of 10 antibodies, only SNF2H, SNF2L, Oct1 and Sox3 have shown promising results determined by Western blotting. However, we cannot completely rule out the possibility that the other antibodies might still recognize the target proteins unless we 1) affinity purify the antibody and 2) test the antibodies in lysate from different zebrafish embryonic stages.

Table B.1: Protein and antigen information

Name	family	Antigen	Full length	Cloning sites	D90 serum
Baf180	SWI/SNF	XP_693735.4	1581 AA	N-term: R146-V185 C-term: A1525-L1564	Not promising
SNF2H	ISWI	NP_001075098.2	1035 AA	N-term: M8-G62	Great
SNF2L	ISWI	NP_001093467.1	1036 AA	N-term: M1-E42	OK
CHD1	CHD	XP_009303375	1778 AA	N-term: M1-S40 C-term: P1649-D1689	Not promising
INO80	INO80	NP_001038584.1	1552 AA	N-term: K281-K320 C-term: E1281-G1320	Not promising
Oct1		NP_571513.1	596 AA	N-term: M1-P44 C-term: G412-A451	Great
Sp1		NP_997827.2	594 AA	N-term: E8-S47 C-term: R521-V560	Not promising
Sox3		NP_001001811.2	300 AA	N-term: K10-N30 C-term: G263-T289	OK
Eomesa		NP_571754.3	661 AA	N-term: N30-T69 C-term: S610-H649	Not promising
Zeb2b		NP_001232895.1	1219 AA	N-term: L64-E103 C-term: N1270-V1219	Not promising

Table B.2: Oligos and primers sequence for antigen peptide

	BAF180
Oligo1	CAGCCGCACATCTTTgcgGCGAGCAACTTCCTTAAGGCTCGACA TAGACTCCTCCTCTGTGGCTCCTCCGAGGTTCTCCTGTGGGTC GTCATCATCATCATCTTCATCATAGTCTCCgCGCTGCACAAACTC GTTTTTGGTCCGtGGATCCgcggcg
Oligo2	CGAGCCTTAAGGAAGTTGCTCGCcgcAAAGATGTGCGGCTGACG AAAGAGCAGGAGAGCCGTCTGCCGTCTCATTGGCTGAAGAGTA AAGGAGCTCACAAGACCATGGCGGACGCTCTGTGGCGcCTGCG cGACCTGtaataaGCGGCCGCattctc
Primer1	cgccgcGGATCCaCGGACCAAAAAC
Primer2	gagaatGCGGCCGCttattaCAGGTCg
	SNF2H
Oligo1	CCATAGCTGTATCTGATGCTTCATCATGGCCATTTTCTGAAGTAT CATCCTCTGCCGGACCGGCCGGGTCGAGTTGGCTCGGCTCTTC TTCCCGCTGCTGCTCCATtGGATCCgcggcg
Oligo2	TCAGAAAATGGCCATGATGAAGCATCAGATACAGCTATGGATAG TGAAGCACAGTCGTCATCACCTCCaTCAGCAAAGGGAAAGGAC CTGCTGGGTGGAtaataaGCGGCCGCattctc
Primer1	cgccgcGGATCCaATGGAGCAG
Primer2	GagaatGCGGCCGCttattaTCCACCCAG
	SNF2L
Oligo1	cgccgcGGATCCaATGTCGGACGAAGAGCTgCCGTCTACTTCTCG GGAGGCGCACGCGGAGGCGGAAGAGGAaGATCCGCCATTTCT CCTGGGTCCTCCTCCTAAAGCAGTGGTGAAGGAGAGCCTGGCT GACCCTGAGtaataaGCGGCCGCattctc
Oligo2	GagaatGCGGCCGCttattaCTCAGGGTCAGCCAGGCTCTCCTTCAC CACTGCTTTAGGAGGAGGACCCAGGAGAAATGGCGGATCtTCC TCTTCCGCCTCCGCGTGCGCCTCCCGAGAAGTAGACGGcAGCT CTTCGTCCGACATtGGATCCgcggcg
	CHD1
Oligo1	GCCGTTGCTGAAGGAGGAATAtGGACTGCTACTGCTTGAACTCG AGCCAGAGCCAGAACCTGATCCTGACCCGGACTCATCATCTGA ATTACTGGACTCTCCGCTGCTGTTGCTGACACTTTCATCCTCGC TGCGTCCATCCATtGGATCCgcggcg
Oligo2	GTTCAAGCAGTAGCAGTCCaTATTCCTCCTTCAGCAACGGCAAA GACCACcgcGAGCGGGAGCGcGACCAGTACcgcGAGcgcGGAGA GCGGCCGGACAGCcgcTACTACAGTGAGAGCAAGCACcgcAAGC TGGACtaataaGCGGCCGCattctc
Primer1	CgccgcGGATCCaATGGATGGAC
Primer2	GagaatGCGGCCGCttattaGTCCAGCTTg

Table B.2: Continued

	INO80
Oligo1	CTTTCCTCCAGCTGgcgTTTCTCTTCTTTGCAGTTCTTCTGTGCCT GGATGGCAGCgcgCCGGACCTCACGCATACACTGGTGGGCCAA CTTgcgAGCATTGGTTAAAACAAAGTTgcgTGCAGAGACTTTTTGT TTAAATGCCTTtGGATCCgcggcg
Oligo2	CGAGCCTTAAGGAAGTTGCTCGCcgcAAAGATGTGCGGCTGACG AAAGAGCAGGAGAGCCGTCTGCCGTCTCATTGGCTGAAGAGTA AAGGAGCTCACAAGACCATGGCGGACGCTCTGTGGCGcCTGCG cGACCTGtaataaGCGGCCGCattctc
Primer1	cgccgcGGATCCaAAGGCATTTAAAC
Primer2	gagaatGCGGCCGCttattaACCTTCCTTC
	Oct1
Oligo1	ACAGCTGAACTTGACTCCAGCTGTGACCTCGCAGACCCCGAGC TCCTCCGCCTCCGCTGGTAAACAGGAAATCCGCGCTGCCACTT CCTCTGCTGCCACACTCACACAGGGGGCGGGGCCATTGCTGG TGACAGGCGCATAAGAGCTCCCCG
Oligo2	CGGGGTCTGCGAGGTCACAGCTGGAGTCAAGTTCAGCTGTGG GAGCAGCGCCTGCACACCTCCGACCGCACCTGACGCTGCACC TGCACTTGGTTTCTGTGCGTCCACACCGTTAGTCTGAGCTTGCG CTGCGCTCTCCATaGGATCCGCCG
Primer1	CGGCGGATCCtATGGAGAGC
Primer2	gagaatGCGGCCGCTTATGCGCCTGTCACCAGCA
	Sp1
Oligo1	GGACTCGCAGCAGCCATCACGCAGTGATCATCTCTCCAAACAC ATCAAGACACACCTTAACAAGAAGGTGTGTGTGAGCGCGTGCA GTGACCCAGTGACCCCGGCTGCAGATGCAGACCCACAGACACT GCTGACGGTGTAAGAGCTCCCCG
Oligo2	TTTGGAGAGATGATCACTGCGTGATGGCTGCTGCGAGTCCTGC GCCGAGCTGGAGCTGCTCTTCTGTGGGAAGCCTGGCTCTGAGT CCGCCATGGAGCTCTTCTGTGGATCGCTCATACGGTGCGATGG CGCGAGTTTTTCtGGATCCGCCG
Primer1	CGGCGGATCCaGAAAAACTCG
Primer2	gagaatGCGGCCGCTTACACCGTCAGCAGTGTCT

Table B.2: Continued

	Sox3
Oligo1	GATCCaAAAAGCCCAATTCCGCAGTCCAACACGGGCTCGGTGA CGGGCGGCAAAAACAACAGTGCCAACGGTGGAGACAGCGCCG ACCACTCCAGTCTGCAGACCAGTCGGTTACACAGCGTTCATCC GCACTATCAAAGCGCAGGGACATAAGAGCT
Oligo2	CTTATGTCCCTGCGCTTTGATAGTGCGGATGAACGCTGTGTAAC CGACTGGTCTGCAGACTGGAGTGGTCGGCGCTGTCTCCACCGT TGGCACTGTTGTTTTTGCCGCCCGTCACCGAGCCCGTGTTGGA CTGCGGAATTGGGCTTTTtG
	Eomesa
Oligo1	TGAGAGCTTTACCGGGACTAGCGGTTACTCCATGGCCTGTAAG CGGCGGCGTGTTTCGCCTGGAGGCTCCAGTACAGACGCTTCAC CAACTATCAAGTGTGAGGACTTGAGCTCTGAGGAGTACAACAAA GAGAGCCATTAAGAGCTCCCCG
Oligo2	ACAGGCCATGGAGTAACCGCTAGTCCCGGTAAAGCTCTCACTG TCAGCATCATCCATCAGAAACTTCTTGGCACCAGAACTCTGCTC CGAATCCGTCCGGTCCATGTCCTGAAAGTCGATCTGCGTTGAC CCCGGGCTATTtGGATCCGCCG
Primer1	CGGCGGATCCaAATAGCCCG
Primer2	gagaatGCGGCCGCTTAATGGCTCTCTTTGTTGT
	Zeb2b
Oligo1	CGAGAACGTTTGGCATGAAAACAAAAGCACGGACGGTGACACT ATGCGCGACGAGGAGGAAAATGGAGAGCACTCCATGGACGAAA GCTCCAAATCGGAGTCCAAATCGGACCATGAGGATGCGATGGA GGACGGGGTGTAAGAGCTCGGGC
Oligo2	GTCACCGTCCGTGCTTTTGTTTTCATGCCAAACGTTCTCGATGC CGCTGTCCTTCATGTCATCTTCATCATCCACAGTCTTGCGCAGC AGTGTGTGACTCAGAGTCGGAGATGGCGGAGCCTCAGACTCAT GATTGACAAGtGGATCCGCCG
Primer1	CGGCGGATCCaCTTGTCAATC
Primer2	GCCCGAGCTCTTACACCC

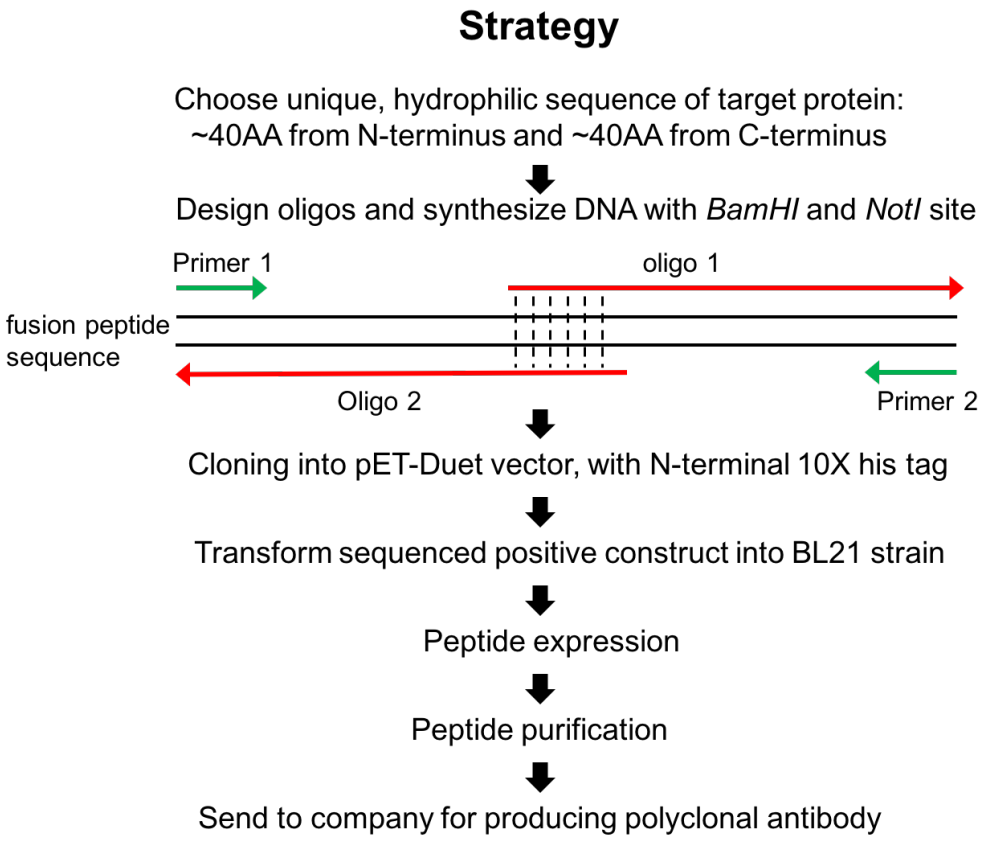


Figure B.1: Strategy for antigen peptide cloning and expression

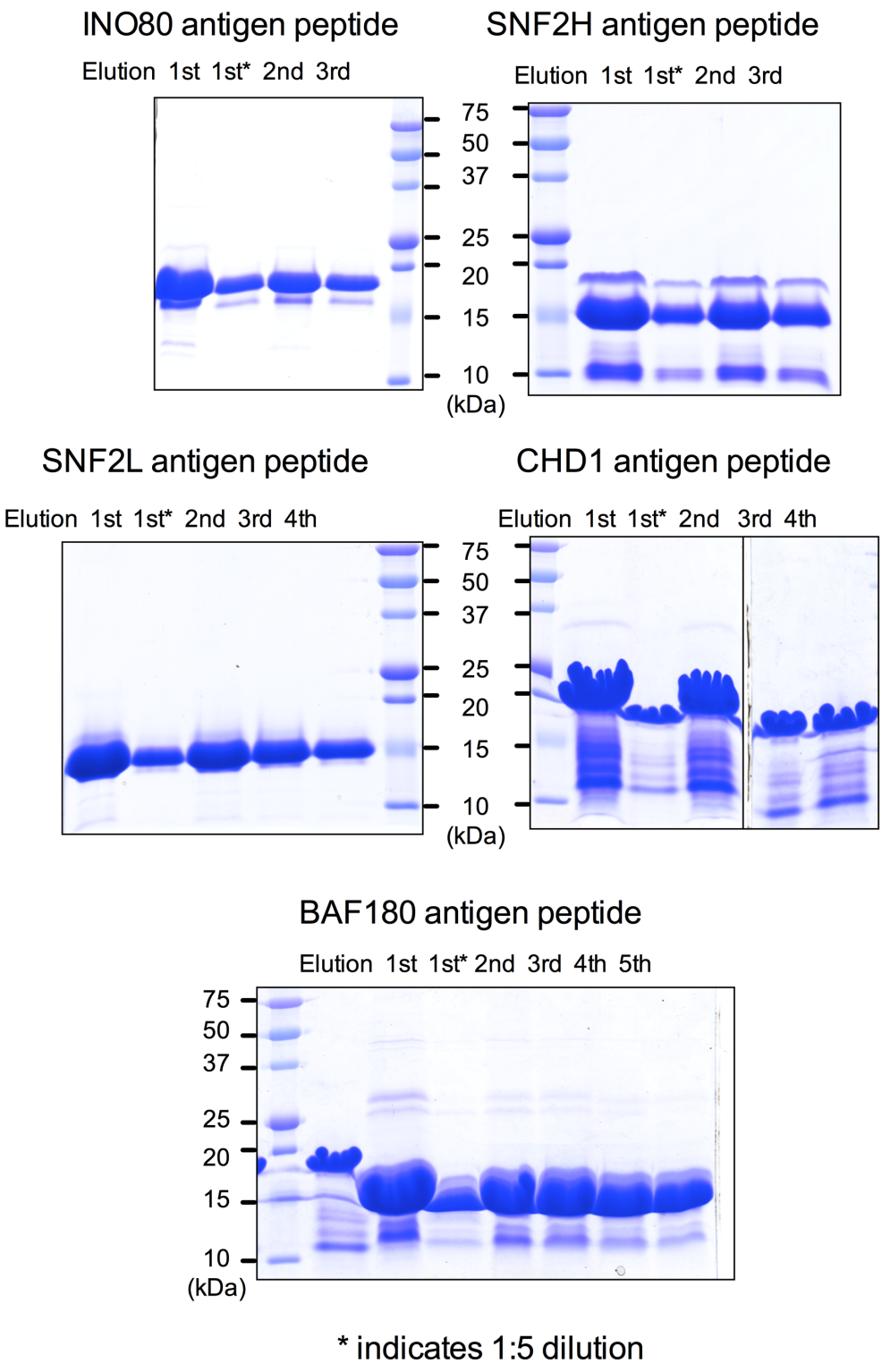


Figure B.2: Coomassie blue staining of purified antigen peptides

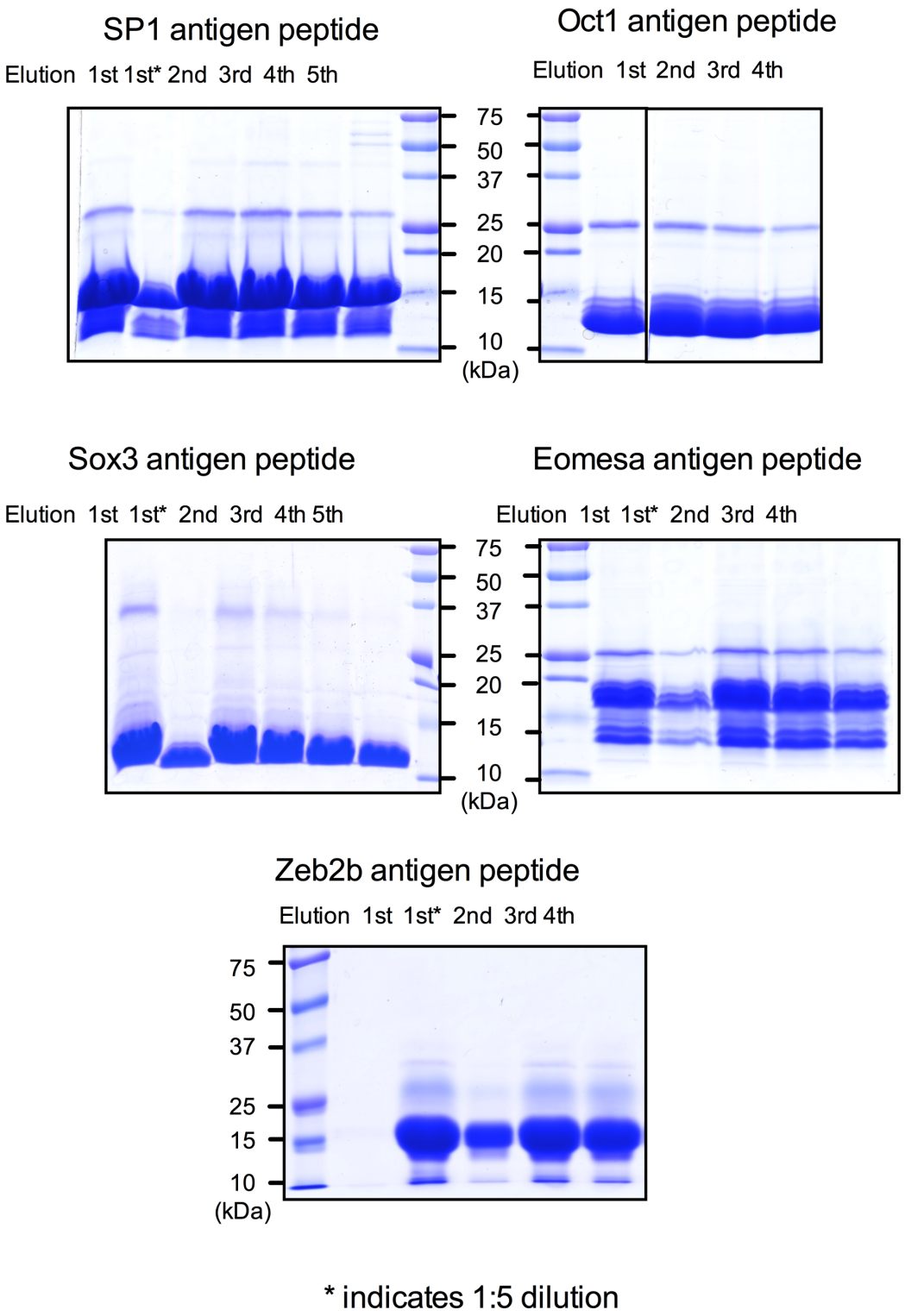


Figure B.2: Continued.

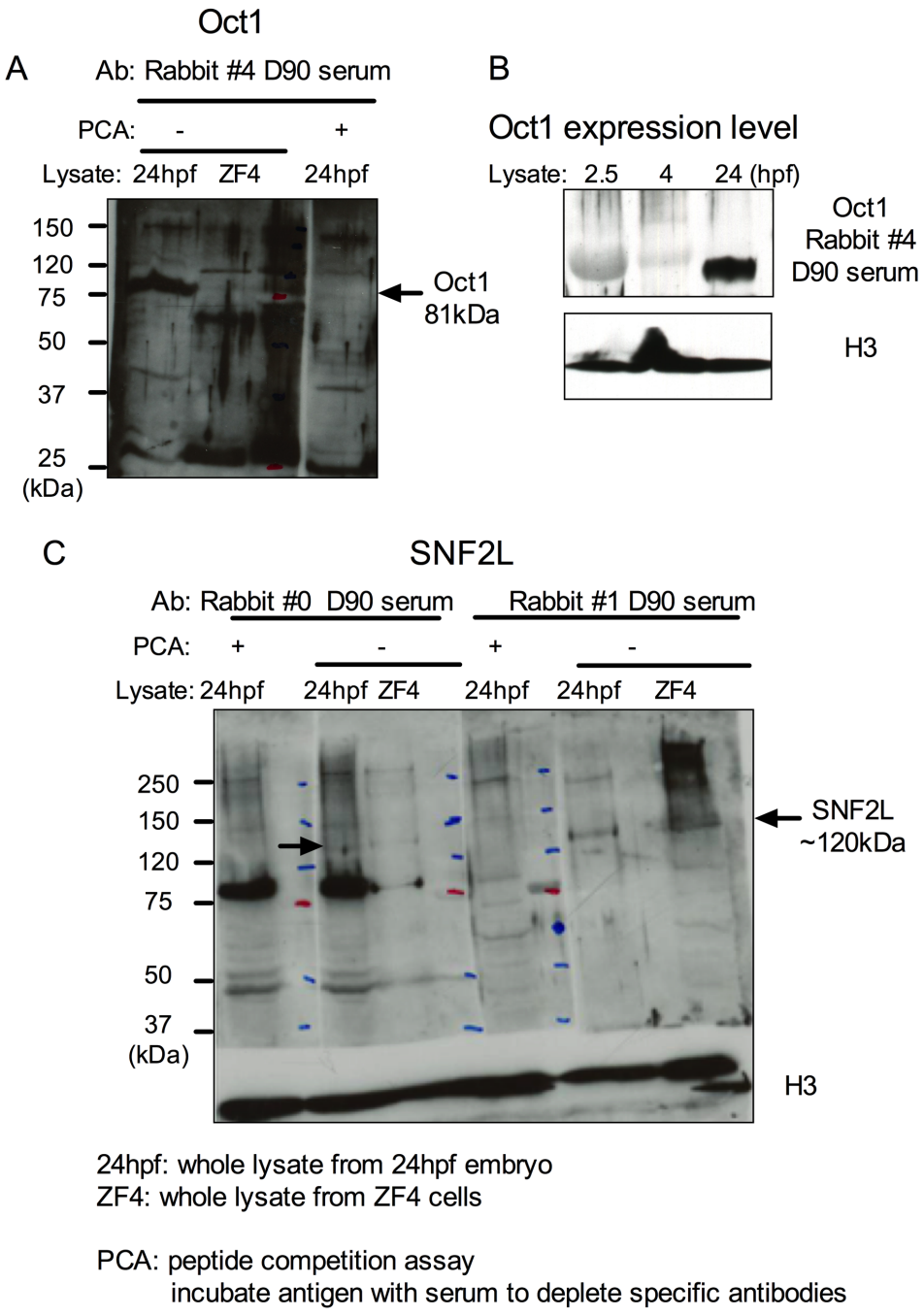
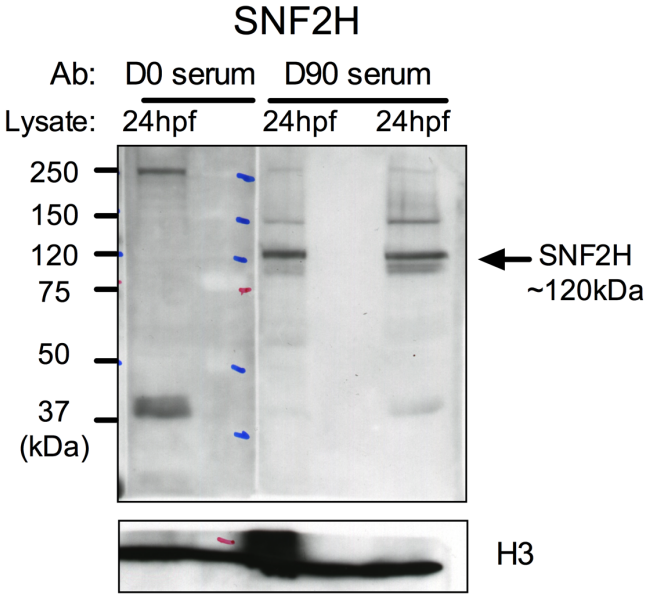


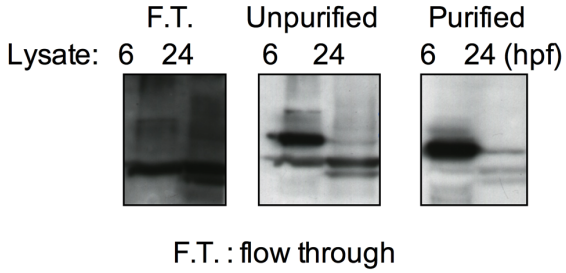
Figure B.3: Western blotting of Oct1 and SNF2L. (A) Western blotting against Oct1 in protein lysate from 24hpf embryo or ZF4 fibroblast cell lines. PCA (peptide competition assay) is to confirm the antibody specificity. (B) Oct1 expression levels in early development. (C) Western blotting against SNF2L in protein lysate from 24hpf embryo or ZF4 fibroblast cell lines, and PCA.

Figure B.4: Western blotting of SNF2H. (A) Western blotting against SNF2H using D0 serum (control) and D90 serum in protein lysate from 24hpf. (B) Western blotting using flow-through, unpurified antibody and purified antibody against SNF2H in 6hpf and 24hpf whole embryo lysate. (C) SNF2H expression levels in early development.

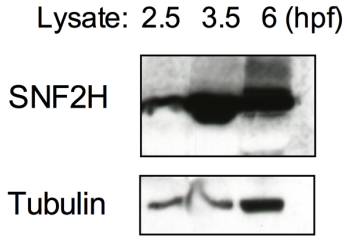
A



B



C



lysate: whole lysate from different embryo stages

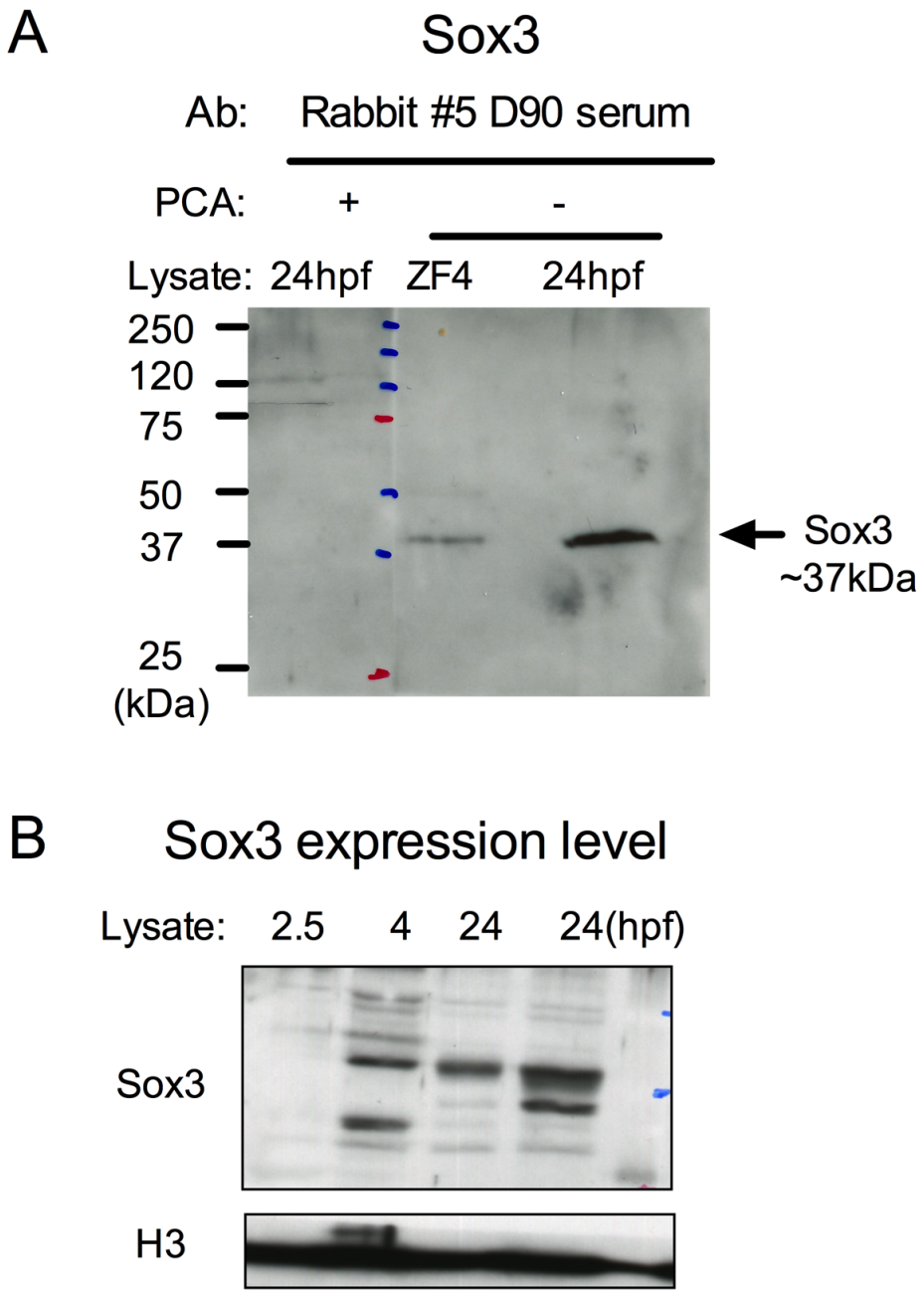


Figure B.5: Western blotting of Sox3. (A) Western blotting against Sox3 in protein lysate from 24hpf embryo or ZF4 fibroblast cell lines, and PCA. (B) Sox3 expression levels in early development.

APPENDIX C

CDNA CONSTRUCTS OF CHROMATIN REMODELERS AND TWO TRANSCRIPTION FACTORS

C.1 Wildtype cDNA construct

I have extensively tested commercial antibodies and customized polyclonal antibodies, in the hope of finding antibodies that are applicable for ChIP-seq. Some candidates can be robustly recognized by their specific antibodies, while others are still lacking specific antibodies. To circumvent the constraint from antibody, I decided to overexpress cDNA tagged by V5. This enables us to investigate the genome-wide occupancy of candidate proteins by performing ChIP-seq with V5 antibody. In summary, we have cloned cDNA for four chromatin remodeler subunits (SNF2H, SNF2L, CHD1 and INO80), and three transcription factors (Pou5f3, Nanog, and Sox19b) (Table C.1.). The cDNAs have been engineered into pCS2+ vector containing 2XV5 tags at the downstream of multicloning sites. Translational efficiency of *in vitro* transcribed RNA has not been investigated.

C.2 Dominant negative construct

Since protein levels of Brg1, Brm and SNF2H are largely unaffected by morpholino injection, I speculate maternally inherited SNF2H protein may be sufficient to support the embryo through the early development. To disrupt the maternal pro-

tein, I cloned catalytic dead constructs for Brg1, Brm and SNF2H. Point mutations for zebrafish Brg1 is K797R, for zebrafish Brm is K757R, and for zebrafish SNF2H is K187R. In principle, the catalytic dead proteins can still form complexes with other subunits. However, they lose the ability to slide or eject nucleosome due to the impaired ATPase function, and thus serve as dominant negative mutant.

Table C.1: Primers for cloning full length cDNA

	Brg1
Brg1-F	ATGTCCACTCCTGACCCACCCATG
Brg1-R	ATCTTCCTCGCTGCCACTAGCC
EcoRI-Brg1-F	cgGAATTCcgccaccATGTCCACTCC
Brg1-XhoI-R	ccgCTCGAGggATCTTCCTCGCTG
	Brg1 K797R (dominant negative)
K797R-F	CTGGCTGATGAAATGGGTTTGGGCAGAACTATCCAGAC CATCGCACTCATC
K797R-R	GATGAGTGCGATGGTCTGGATAGTTCTGCCCAAACCCA TTTCATCAGCCAG
	Brm
Brm-F	ATGGCCACCCCCAATGAGC
Brm-R	CTCATCATCGCTCTGGCTTCCTTC
ClaI-Brm-F	ccATCGATggccaccATGGCCACC
Brm-SpeI-R	ggactagtccCTCATCATCGCTCTGGCTTCC
	Brm K757R (dominant negative)
K757R-F	CTGGCGGACGAGATGGGTCTGGGGAGGACCATTCAAA CTATCGGGCTTATC
K757R-R	GATAAGCCCGATAGTTTGAATGGTCCTCCCCAGACCCA TCTCGTCCGCCAG
	SNF2H
SNF2H-F	ATGGAGCAGCAGCGGGAAGAAG
SNF2H-R	CAATTTCAGCTTTTTCTTGCGGCCACG
EcoRI-Snf2H-F	CGGAATTCCgccaccATGGAGCAGC
snf2H-SpeI-R	ggactagtccCAATTTCAGCTTTTTCTTGCGGCC
	SNF2H K187R (dominant negative)
K187R-F	CTTGCTGATGAAATGGGTCTTGGTAGGACCCTCCAGAC CATCTCTCTGCTG
K187R-R	CAGCAGAGAGATGGTCTGGAGGGTCCTACCAAGACCCA TTTCATCAGCAAG
	SNF2L
SNF2L-F	ATGTCGGACGAAGAGCTACCGTCTAC
SNF2L-R	GGTACGAGCTTTCTTCTCCTTTTTCTCAGATGAC
EcoRI-Snf2L-F	CGGAATTCCgccaccATGTCGGAC
snf2L-SpeI-R	ggactagtccGGTACGAGCTTTCTTCTCC
	CHD1
CHD1-F	ATGGATGGACGCAGCGAGGATG
CHD1-R	TGTCTTGCGGCCGCTCC
BamHI-CHD1-F	cgGGATCCCgccaccATGGATGG
CHD1-SpeI-R	ggactagtccTGTCTTGCGGCCGC

Table C.1: Continued

	INO80
INO80-F	ATGGCCTCAGGGCAGGATGG
INO80-R	CCGTCCACTGTGGTTGCTGG
EcoRI-INO80-F	cgGAATTCCgccaccATGGCCTCAG
INO80-SpeI-R	ggactagtccCCGTCCACTGTGGTTG
	Oct4
SNF2L-F	ATGACGGAGAGAGCGCAGAGC
SNF2L-R	GCTGGTGAGATGACCCACCAAACC
EcoRI-Snf2L-F	cgGAATTCcgccaccATGACGGAGAGAG
snf2L-SpeI-R	ggactagtccGCTGGTGAGATGACCCACC
	Nanog
Nanog-F	ATGGCGGACTGGAAGATGCCAG
Nanog-R	CAGCAAAGTTATTCCTTTAGTTGCCCACAGG
EcoRI-Nanog-F	cgGAATTCcgccaccATGGCGGAC
Nanog-SpeI-R	ggactagtccCAGCAAAGTTATTCCTTTAGTTGCCCAC
	Sox19b
Sox19b-F	ATGATGTACAGCATGATGGAGCACGAG
Sox19b-R	GATGTGAGTGAGGGGAACAGTTCC
EcoRI-Sox19b-F	cgGAATTCcgccaccATGATGTACAGCATGATGG
Sox19b-SpeI-R	ggactagtccGATGTGAGTGAGGGGAACAGTTCC